

Técnicas de Análisis Multivariante de Datos

Aplicaciones con SPSS®



César Pérez

Técnicas de Análisis Multivariante de Datos

Aplicaciones con SPSS®

CÉSAR PÉREZ LÓPEZ

*Universidad Complutense de Madrid
Instituto de Estudios Fiscales*



Madrid • México • Santafé de Bogotá • Buenos Aires • Caracas • Lima
Montevideo • San Juan • San José • Santiago • São Paulo • White Plains

Datos de catalogación bibliográfica

CÉSAR PÉREZ LÓPEZ

Técnicas de Análisis Multivariante de Datos

PEARSON EDUCACIÓN, S.A., Madrid, 2004

ISBN: 978-84-205-4104-4

MATERIA: Estadística Matemática 519.2

Formato: 170 × 240 mm

Páginas: 672

Todos los derechos reservados. Queda prohibida, salvo excepción prevista en la ley , cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de la propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. Código Penal*).

DERECHOS RESERVADOS

© 2004 por PEARSON EDUCACIÓN, S.A.
Ribera del Loira, 28
28042 Madrid (España)

PEARSON PRENTICE HALL es un sello editorial autorizado de PEARSON EDUCACIÓN, S.A.

CÉSAR PÉREZ LÓPEZ

TÉCNICAS DE ANÁLISIS MULTIVARIANTE DE DATOS

ISBN: 978-84-205-4104-4

Depósito Legal: M. 39.243-2008
Última reimpresión, 2008

Equipo editorial

Editor: Isabel Capella

Técnico editorial: Marta Caicoya

Equipo de producción:

Director: José A. Clares

Técnico: Diego Marín

Diseño de cubierta: Equipo de diseño de PEARSON EDUCACIÓN, S.A.

Impreso por: Gráficas Rógar, S.A.

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

Este libro ha sido impreso con papel y tintas ecológicos

ÍNDICE

<i>Introducción</i>	<i>XV</i>
<i>Capítulo 1. Introducción a las técnicas de análisis multivariante de datos</i>	<i>1</i>
Clasificación de las técnicas de análisis multivariante de datos por objetivo principal.....	1
Clasificación de las técnicas de análisis multivariante de datos por tipo de variables.....	2
Clasificación global de las técnicas de análisis multivariante de datos.....	3
Métodos explicativos: Técnicas del análisis de la dependencia.....	4
Regresión múltiple	5
Análisis canónico (correlación canónica).....	5
Análisis discriminante	5
Modelos de elección discreta	6
Modelo ANOVA (Análisis de la varianza simple).....	7
Modelo ANCOVA (Análisis de la covarianza simple)	7
Modelo MANOVA (Análisis de la varianza múltiple)	8
Modelo MANCOVA (Análisis de la covarianza múltiple)	8
Análisis conjunto	9
Segmentación jerárquica	9
Regresión múltiple y modelos de elección discreta con variables ficticias..	10
Métodos descriptivos: Técnicas del análisis de la interdependencia.....	11
Análisis en componentes principales.....	11
Análisis factorial.....	12
Análisis de correspondencias	13
Análisis de conglomerados (análisis cluster)	14
Escalamiento multidimensional.....	15
Técnicas emergentes de análisis multivariante de datos	16
Fases a seguir en las técnicas de análisis multivariante de datos	18

Capítulo 2. Primeros pasos en el análisis multivariante. Análisis exploratorio de datos	21
Análisis previo de los datos.....	21
Análisis exploratorio y gráfico de los datos	22
Histograma de frecuencias	23
Diagrama de tallo y hojas	25
Gráfico de caja y bigotes	26
Gráfico múltiple de caja y bigotes.....	28
Gráfico de simetría	30
Gráfico de dispersión.....	32
Estadísticos robustos	33
Transformaciones de las variables.....	39
Análisis de los datos ausentes	39
Detección y diagnóstico de los datos ausentes	40
Soluciones para los datos ausentes: Supresión de datos o imputación de la información faltante	46
Análisis y detección de valores atípicos.....	48
Detección univariante de casos atípicos	49
Detección bivariante de casos atípicos	54
Detección multivariante de casos atípicos.....	55
Comprobación de los supuestos del análisis multivariante	55
Normalidad.....	56
Heteroscedasticidad.....	63
Multicolinealidad.....	65
Autocorrelación	65
Linealidad.....	66
Análisis de los residuos	68
Capítulo 3. SPSS y el análisis exploratorio de datos. Datos atípicos y ausentes	69
Análisis exploratorio de los datos con SPSS. El procedimiento Explorar	69
Gráficos de análisis exploratorio con SPSS	74
Tipos de gráficos	74
Histogramas.....	75
Gráficos de normalidad	75
Gráficos de caja y bigotes	78
Gráficos de control para la detección de datos atípicos	80
Gráficos de dispersión	82
Tratamiento de los valores ausentes con SPSS	84
Diagnóstico para los datos ausentes con SPSS. El procedimiento Prueba T para muestras independientes	84
Diagnóstico de los datos ausentes con SPSS: El procedimiento Correlaciones bivariadas	86
Técnicas de imputación de datos ausentes con SPSS. El procedimiento Reemplazar valores perdidos.....	88
Supresión de los datos ausentes con SPSS	89

El procedimiento frecuencias de SPSS.....	90
El procedimiento descriptivos de SPSS	92
Los procedimientos informe de estadísticos en filas y columnas de SPSS	94
El procedimiento Resumir de SPSS	96
Ejercicio 3-1	99
Ejercicio 3-2	107
Capítulo 4. Análisis en componentes principales	121
Objetivo del análisis en componentes principales.....	121
Obtención de las componentes principales.....	123
Varianzas de las componentes.....	127
Estructura factorial de las componentes principales	128
Puntuaciones o medición de las componentes.....	129
Contrastes sobre el número de componentes principales a retener	130
Criterio de la media aritmética	130
Contraste sobre las raíces características no retenidas	131
Prueba de Anderson.....	132
Prueba de Lebart y Fenelón.....	132
Prueba del bastón roto de Frontier	133
Prueba □ de Ibáñez.....	134
El gráfico de sedimentación	134
Retención de variables.....	134
La regresión sobre componentes principales y el problema de la multicolinealidad	135
La regresión ortogonal y las componentes principales.....	137
Interpretación geométrica del análisis en componentes principales.....	138
El hiperelipsoide de concentración.....	141
Matriz de cargas factoriales, communalidad y círculos de correlación.....	144
Rotación de las componentes	145
El caso de dos variables.....	147
Propiedades muestrales de las componentes principales.....	153
Capítulo 5. Análisis factorial	155
Objetivo del análisis factorial	155
El modelo factorial	158
Hipótesis en el modelo factorial	158
Comunalidades y especificidades.....	159
Método de Turstone para obtener los factores	160
Método del factor principal para obtener los factores	162
Método Alpha para obtener los factores.....	165
Método del centroide para obtener los factores.....	165
Método de las componentes principales para obtener los factores.....	167
Método de componentes principales iteradas o ejes principales para obtener los factores	169
Método de máxima verosimilitud para obtener los factores	170
Métodos Minres, ULS y GLS para obtener los factores	173

Contrastes en el modelo factorial	175
Contraste de esfericidad de Barlett.....	175
Medida KMO de Kaiser, Meyer y Olkin de adecuación muestral global al modelo factorial y medida MSA de adecuación individual	176
Contraste de la bondad de ajuste en el método ML de máxima verosimilitud.....	177
Contraste de la bondad de ajuste en el método MINRES	179
Interpretación geométrica del análisis factorial.....	179
Rotación de los factores	182
Rotaciones ortogonales.....	183
Método Varimax	183
Método Quartimax	184
Métodos Ortomax: Ortomax general, Biquartimax y Equamax.....	186
Rotaciones oblicuas	186
Método Oblimax y método Quartimin	186
Métodos Oblimin: Covarimin, Oblimin general y Biquartimin	187
Método Oblimin Directo: Rotación Promax.....	188
Puntuaciones o medición de los factores.....	189
Medición de componentes principales	189
Medición de factores mediante estimación por mínimos cuadrados.....	190
Medición de los factores mediante estimación por regresión	190
Medición de los factores mediante el método de Barlett	191
Medición de los factores mediante el método de Anderson y Rubin	191
Análisis factorial exploratorio y confirmatorio	191
Capítulo 6. Componentes principales y análisis factorial con SPSS.....	193
Componentes principales y análisis factorial	193
Esquema general del análisis factorial	194
SPSS y el análisis factorial	195
Ejercicio 6-1	203
Ejercicio 6-2	212
Capítulo 7. Métodos factoriales en general. Análisis de correspondencias ..	219
Cantidad de información y distancias.....	219
Análisis general de los métodos factoriales.....	222
Objetivo general del análisis factorial	223
Análisis en R^P	223
Análisis en R^n	227
Relaciones entre los análisis en R^P y R^n	228
Reconstrucción de la tabla inicial de datos a partir de los ejes factoriales...	230
Componentes principales como caso particular del análisis factorial general ..	231
Análisis en R^P	231
Análisis en R^n	234
Análisis factorial de correspondencias	236
Análisis de correspondencias simples	237
Formación de las nubes y definición de distancias	239
Ejes factoriales: Análisis en R^P	241

Ejes factoriales: Análisis en R^n	244
Relaciones entre los análisis en R^p y R^n	244
Reconstrucción de la tabla de frecuencias.....	245
Análisis de correspondencias múltiples.....	246
Obtención de los factores: Tabla de Burt	248
Capítulo 8. SPSS y el análisis de correspondencias.....	251
SPSS y correspondencias simples	251
SPSS y las correspondencias múltiples	260
Aplicaciones del análisis de correspondencias.....	266
Ejercicio 8-1	268
Ejercicio 8-2	271
Capítulo 9. Escalamiento óptimo y multidimensional	275
Concepto de escalamiento óptimo.....	275
Correlación canónica no lineal	278
Análisis de componentes principales categóricas.....	280
Concepto de escalamiento multidimensional	281
Mapas perceptuales	283
Solución, ajuste y preferencias en el escalamiento multidimensional.....	285
Interpretación y validación de los resultados	289
Modelos de escalamiento multidimensional	291
Modelo de escalamiento métrico.....	292
Modelo de escalamiento no métrico.....	294
Modelo de escalamiento de diferencias individuales	295
Modelos de escalamiento para datos de preferencia	297
Modelos de escalamiento desdoblado (<i>unfolding</i>)	297
Modelo de escalamiento vectorial	298
Interpretación de los resultados obtenidos	299
Interpretación dimensional	299
Interpretación por agrupamientos.....	300
Interpretación por regiones.....	301
Aplicaciones del MDS y su relación con otras técnicas de análisis de datos...	301
Capítulo 10. Escalamiento óptimo y multidimensional en SPSS.....	303
SPSS y el escalamiento óptimo	303
Análisis de componentes principales categóricas con SPSS	304
Correlación canónica no lineal con SPSS	317
SPSS y el escalamiento multidimensional	324
Procedimiento ALSCAL	324
Procedimiento PROXSCAL	332
Ejercicio 10-1	335
Ejercicio 10-2	339
Ejercicio 10-3	343
Ejercicio 10-4	348
Ejercicio 10-5	353

Capítulo 11. Modelos logarítmico lineales y tablas de contingencia	357
Tablas de contingencia	357
Distribuciones marginales y condicionadas	357
Independencia y asociación de variables cualitativas. Coeficientes	359
El modelo logarítmico lineal	366
Efectos principales	368
Interacciones.....	371
Modelo saturado	372
Modelo jerárquico	372
Independencia y asociación en modelo logarítmico lineales	373
Independencia total, completa o global	373
Independencia parcial: un factor completamente independiente de los demás....	373
Independencia condicional	373
Asociación parcial	373
Estimación máximo verosímil de los parámetros del modelo.....	374
Tablas de contingencia bidimensionales	374
Estimación de la media general.....	376
Estimación del efecto principal del factor A	376
Estimación del efecto principal del factor B	376
Análisis de los residuos	378
Análisis de los residuos de los parámetros del modelo	378
Análisis de los residuos en las celdas	379
Capítulo 12. Modelos logarítmico lineales y tablas de contingencia con SPSS ...	381
El procedimiento Tablas de contingencia.....	381
El procedimiento Resumir.....	385
SPSS y los modelos logarítmico lineales	388
Selección del modelo.....	388
Análisis loglineal general	388
Análisis logit.....	388
Ejercicio 12-1	389
Ejercicio 12-2	398
Ejercicio 12-3	406
Ejercicio 12-4	411
Capítulo 13. Clasificación y segmentación mediante análisis cluster	417
Concepto de análisis cluster	417
Distancias y similitudes.....	419
Clusters no jerárquicos	423
Clusters jerárquicos: Dendograma	427
Fórmula de Lance y Williams para la distancia entre grupos	431
Análisis de conglomerados en dos fases	433

Capítulo 14. Clasificación y segmentación mediante análisis cluster con SPSS.....	435
Principios del análisis cluster	435
Esquema general del análisis cluster	435
SPSS y el análisis cluster no jerárquico	437
SPSS y el análisis cluster jerárquico	440
SPSS y el análisis cluster en dos fases	445
Consideraciones previas	446
Ejercicio 14-1	447
Ejercicio 14-2	452
Capítulo 15. Clasificación y segmentación mediante análisis discriminante	457
Concepto de análisis discriminante	457
Clasificación con dos grupos.....	458
Contrastes y probabilidad de pertenencia (2 grupos)	463
Clasificación con más de dos grupos	470
Análisis discriminante canónico.....	473
Capítulo 16. SPSS y la Clasificación y segmentación mediante análisis discriminante	475
Principios del análisis discriminante	475
Esquema general del análisis discriminante	476
SPSS y el análisis discriminante.....	477
Ejercicio 16-1	483
Ejercicio 16-2	487
Capítulo 17. Análisis de la varianza y la covarianza	495
Introducción al análisis de la varianza	495
Análisis de la varianza simple (un solo factor): modelo unifactorial de efectos fijos.....	497
Modelo unifactorial de efectos aleatorios.....	501
Análisis de la varianza con varios factores: modelo bifactorial de efectos fijos ANOVA IIF.....	503
Modelo bifactorial general con efectos aleatorios ANOVA IIA.....	508
Modelo bifactorial general con efectos mixtos ANOVA IIM.....	509
Modelo en bloque aleatorizados.....	510
Modelo ANOVA factorial con tres factores.....	512
Modelo en cuadrado latino	513
Modelos de la covarianza ANCOVA	514
Modelo con un factor y un covariante	514
Modelos con dos factores y un covariante	514
Modelo con dos factores y dos covariantes	515
Modelo MANOVA (Análisis de la varianza múltiple)	515

Modelo MANCOVA (Análisis de la covarianza múltiple)	516
Modelo lineal general (GLM)	516
Capítulo 18. Análisis de la varianza y la covarianza con SPSS.....	517
Procedimiento ANOVA de un factor	517
Procedimiento MLG univariante.....	520
Procedimiento MLG multivariante.....	527
Procedimiento MLG medidas repetidas	530
Procedimiento Componentes de la varianza.....	533
Ejercicio 18-1	535
Ejercicio 18-2	538
Ejercicio 18-3	540
Ejercicio 18-4	542
Ejercicio 18-5	545
Ejercicio 18-6	547
Capítulo 19. Modelos de elección discreta Logit y Probit. Regresión de Cox.....	551
Modelos con variables cualitativas: modelos de elección discreta	551
El modelo de regresión logística	553
Interpretación de los coeficientes del modelo	555
Estimación de los coeficientes.....	556
Estimación por intervalos y contrastes de hipótesis sobre los coeficientes.....	557
Modelos Probit y Logit.....	558
Análisis de la supervivencia	559
Tablas de mortalidad	560
Estimaciones no paramétricas de la función de supervivencia	561
Estimaciones paramétricas de la función de supervivencia	561
Regresión de Cox	562
Capítulo 20. SPSS y los modelos de elección discreta Logit y Probit. Regresión de Cox	563
SPSS y la regresión logística	563
SPSS y la regresión logística multinomial	569
SPSS y los modelos Logit y Probit	573
Procedimiento Tablas de mortalidad	577
Procedimiento Kaplan-Meier	580
Procedimiento Regresión de Cox	583
Ejercicio 20-1	587
Ejercicio 20-2	588
Ejercicio 20-3	589
Ejercicio 20-4	590

<i>Capítulo 21. Análisis conjunto.....</i>	595
Concepto de Análisis Conjunto	595
El Análisis Conjunto como una técnica multivariante de la dependencia.....	597
Técnicas composicionales y descomposicionales	598
Aplicaciones del Análisis Conjunto	600
Análisis Conjunto a través del perfil completo	603
Análisis Conjunto y diseños de experimentos.....	604
<i>Capítulo 22. SPSS y el análisis conjunto</i>	605
Análisis Conjunto a través de SPSS	605
SPSS y el método del concepto completo	606
Un ejemplo completo a través de SPSS	607
El procedimiento Generar diseño ortogonal.....	609
Configuración del número de tarjetas de estímulos a generar.....	612
Preparación de tarjetas de estímulos.....	613
Recogida de los datos	621
Análisis de las preferencias mediante Análisis Conjunto.....	621
Interpretación de las salidas del Análisis Conjunto.....	624
Ejercicio 22-1	629
Ejercicio 22-2	631
<i>Índice alfabético</i>	635

INTRODUCCIÓN

El análisis estadístico de datos incluye un conjunto de *métodos y técnicas univariantes y multivariantes* que permiten estudiar y tratar en bloque una o varias variables medidas u observadas en una colección de individuos. Existe la posibilidad de que estas variables sean sólo cuantitativas, sólo cualitativas, o simultáneamente de ambos tipos. Un tratamiento tan completo, unido a la diversidad de enfoques teóricos y prácticos que puede darse a un estudio multidimensional, explica la dificultad matemática de un proceso que, por fuerza, ha de apoyarse en el cálculo matricial y en técnicas no básicas. Es ésta la razón por la cual, hasta época muy reciente, no ha comenzado a difundirse su aplicación con la frecuencia necesaria para que la investigación científica se beneficie del empleo de técnicas tan avanzadas.

Ha sido preciso un espectacular desarrollo del proceso automático de datos, hoy al alcance de cualquier equipo de trabajo, para asistir a una generalización del uso de los métodos de análisis multivariante de datos a través de potentes programas específicos de computador que, a una gran capacidad de manejo de información, añaden una alta velocidad de proceso, a la vez que una relativa facilidad de utilización por grupos de investigadores que no tienen por qué ser expertos matemáticos o informáticos.

La exigencia de cálculos tediosos que sobrepasaban la capacidad y la velocidad de los computadores habituales existentes en la época del desarrollo de estos métodos, la inaccesibilidad a los supercomputadores, entonces restringidos a grandes compañías o instituciones, y la complejidad de su manejo, limitada a unos pocos especialistas por la no existencia de aplicaciones específicamente diseñadas para estas técnicas, hacían impensable su difusión.

Por otra parte, la no disponibilidad de textos comprensibles orientados fundamentalmente al conocimiento conceptual de los métodos con el mínimo contenido matemático posible y dirigidos a la aplicación práctica explican, sin duda, la tardanza en la utilización de las técnicas multivariantes en el mundo práctico.

No obstante, ya desde los años 70, los métodos de análisis multivariante de datos (sobre todo los métodos factoriales y de clasificación) han probado su eficacia en el estudio de grandes masas de información compleja. Se trata de métodos descriptivos pero a la vez multidimensionales, por oposición a los métodos de estadística descriptiva que tratan únicamente una o dos variables.

Los métodos factoriales, muy utilizados en *Data Mining* (o Minería de datos), permiten confrontar numerosas informaciones, lo que hace que sea más rico que los análisis separados. Ofrecen representaciones gráficas de la información contenida en grandes tablas de datos y se manifiestan así como un instrumento de síntesis notable, mediante una reducción de la dimensionalidad. Permiten extraer las tendencias más destacadas, jerarquizarlas y eliminar los efectos marginales o puntuales que perturban la percepción global de lo observado.

Los métodos de clasificación resultan ser un complemento necesario de los métodos factoriales ya que facilitan la construcción de tipologías de individuos según las variables en estudio.

En los últimos años, y debido en parte a la proliferación de los potentes computadores personales, se ha podido detectar un notable incremento en la aparición de artículos que utilizan técnicas de análisis multivariante de datos en las principales revistas científicas de ámbitos muy dispares, por ejemplo, en el ámbito médico, en el ámbito biológico, en el ámbito de la investigación comercial, en el ámbito de la prospección de mercados, en el ámbito de la minería de datos, etc.

Estos métodos también se aplican en múltiples dominios científicos como la sociología, la epidemiología, la ecología, la lingüística, la psicometría, el análisis de mercados, la arqueología, en la banca y los seguros y en la mayoría de las situaciones en que deban analizarse grandes ficheros de datos.

El objetivo de este libro es proporcionar una visión clara conceptual de las técnicas estadísticas univariantes y multivariantes de análisis de datos, describir pormenorizadamente sus principales métodos (análisis en componentes principales, análisis factorial, análisis discriminante, análisis de correspondencias, análisis cluster, análisis de la varianza, análisis de la regresión múltiple, regresión logística, escalamiento multidimensional, análisis conjunto, etc.) e ilustrar con ejemplos prácticos su aplicación en los distintos campos de la investigación.

Se trata de enseñar al investigador a plantear por sí mismo, y quizá a resolver, problemas habituales, pero, desde luego, se pretende hacerle llegar, con el mínimo esfuerzo por su parte, a dialogar de tú a tú con el análisis multivariante.

Evidentemente esto no se consigue sin el uso de medios informáticos. Por ello, adicionalmente, en este libro se utiliza uno de los programas de computador más usuales en este campo, concretamente SPSS, y se resuelven los problemas con este software.

SPSS es un paquete estadístico de análisis de datos con más de 20 años de aplicación en la investigación de las ciencias sociales y económicas. Contiene programas capaces de realizar desde un simple análisis descriptivo hasta diferentes tipos de análisis multivariante de datos, ya citados anteriormente.

INTRODUCCIÓN A LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE DE DATOS

CLASIFICACIÓN DE LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE DE DATOS POR OBJETIVO PRINCIPAL

Al enfrentarse a la realidad de un estudio, el investigador dispone habitualmente de muchas variables medidas u observadas en una colección de individuos y pretende estudiarlas conjuntamente, para lo cual suele acudir al análisis estadístico de datos univariante y multivariante. Entonces se encuentra frente a una diversidad de técnicas y debe seleccionar la más adecuada a sus datos pero, sobre todo, a su objetivo científico.

Al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los **métodos multivariantes de reducción de la dimensión** (análisis en componentes principales, factorial, correspondencias, escalamiento óptimo, homogeneidades, análisis conjunto, etc.) tratan de eliminarla. Estos métodos combinan muchas variables observadas para obtener pocas variables ficticias que las representen.

Por otro lado, los individuos pueden presentar ciertas características comunes en sus respuestas, que permitan intentar su **clasificación en grupos de cierta homogeneidad**. Los métodos de clasificación (análisis cluster, análisis discriminante, árboles de decisión, etc.) buscan analizar las relaciones entre variables para ver si se pueden separar los individuos en agrupaciones a posteriori.

Finalmente, podrá existir una variable cuya dependencia de un conjunto de otras sea interesante detectar para analizar su **relación** o, incluso, aventurar su **predicción** cuando las demás sean conocidas. En este apartado cabe incluir la regresión lineal simple y múltiple, regresión no lineal, regresión logística, análisis de la varianza simple y múltiple, las técnicas de análisis de series temporales, etc.

CLASIFICACIÓN DE LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE DE DATOS POR TIPO DE VARIABLES

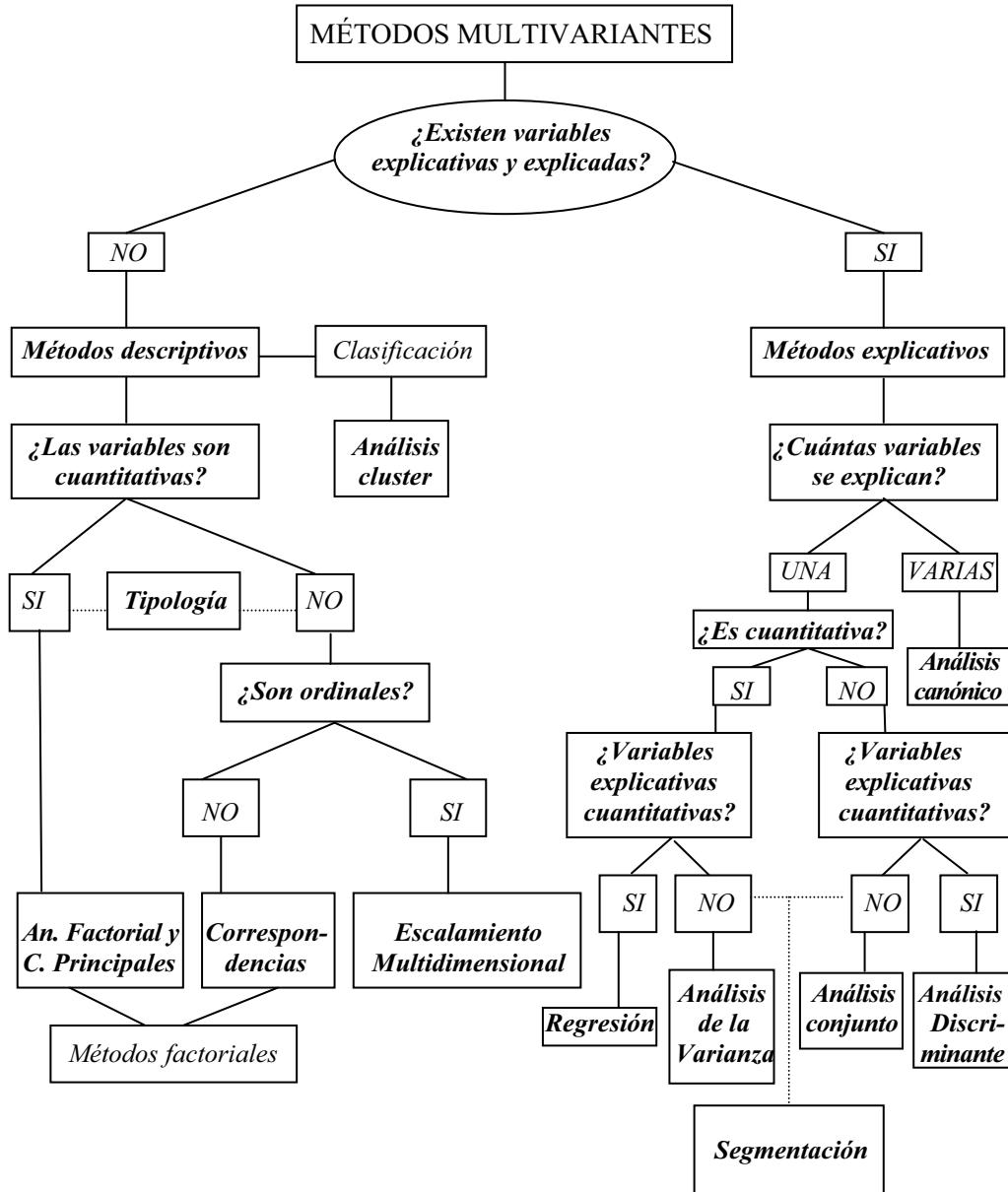
Por otra parte, el investigador tendrá que considerar si asigna a todas sus variables una importancia equivalente, es decir, si **ninguna variable destaca como dependiente principal (MÉTODOS DE INTERDEPENDENCIA)** en el objetivo de la investigación. Si es así, porque maneja simplemente un conjunto de diversos aspectos observados y coleccionados en su muestra, puede acudir para su tratamiento en bloque a lo que podría llamarse **técnicas multivariantes descriptivas**.

Y puede hacerlo con dos orientaciones diferentes: Por una parte, para **reducir la dimensión de una tabla de datos excesivamente grande** por el elevado número de variables que contiene y quedarse con unas cuantas variables ficticias que, aunque no observadas, sean combinación de las reales y sinteticen la mayor parte de la información contenida en sus datos. En este caso también deberá tener en cuenta el tipo de variables que maneja. Si son **variables cuantitativas**, las técnicas que le permiten este tratamiento pueden ser el *Análisis de componentes principales* y el *Análisis factorial*, si son **variables cualitativas**, puede acudir al *Análisis de correspondencias* y si son variables cualitativas ordinales se acude al *Escalamiento multidimensional*. La *Tipología* acepta **variables cualitativas y cuantitativas**. Por otra parte, la otra orientación posible ante una colección de variables sin ninguna destacada en dependencia, sería la de **clasificar sus individuos en grupos más o menos homogéneos en relación al perfil** que en aquéllas presenten, en cuyo caso utilizará por ejemplo el *Análisis de clusters*, donde los grupos, no definidos previamente, serán configurados por las propias variables que utiliza.

Si no fuera aceptable una importancia equivalente en las variables, porque **alguna variable se destaca como dependiente principal (MÉTODOS DE DEPENDENCIA)**, habrá de utilizar **técnicas multivariantes analíticas o inferenciales** considerando la variable dependiente como explicada por las demás variables independientes explicativas, y tratando de **relacionar todas las variables** por medio de una posible ecuación o modelo que las ligue. El método elegido podría ser entonces la *Regresión lineal*, generalmente con todas las variables cuantitativas. Una vez configurado el modelo matemático se podrá llegar a **predecir el valor de la variable dependiente** conocido el perfil de todas las demás. Si la variable dependiente fuera cualitativa dicotómica (1,0; sí o no) podrá usarse como clasificadora, estudiando su relación con el resto de variables clasificativas a través de la *Regresión logística*. Si la variable dependiente cualitativa observada constatará la asignación de cada individuo a grupos previamente definidos (dos, o más de dos), puede ser utilizada para **clasificar nuevos casos en que se desconozca el grupo a que probablemente pertenezcan**, en cuyo caso estamos ante el *Análisis discriminante*, que resuelve el problema de asignación en función de un perfil cuantitativo de variables clasificativas. Si la variable dependiente es cuantitativa y las explicativas son cualitativas estamos ante los *modelos del análisis de la varianza*, que puede extenderse a los *modelos loglineales* para el análisis de tablas de contingencia de dimensión elevada. Si la variable dependiente puede ser cualitativa o cuantitativa y las independientes cualitativas, estamos ante el caso de la *Segmentación*.

CLASIFICACIÓN GLOBAL DE LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE DE DATOS

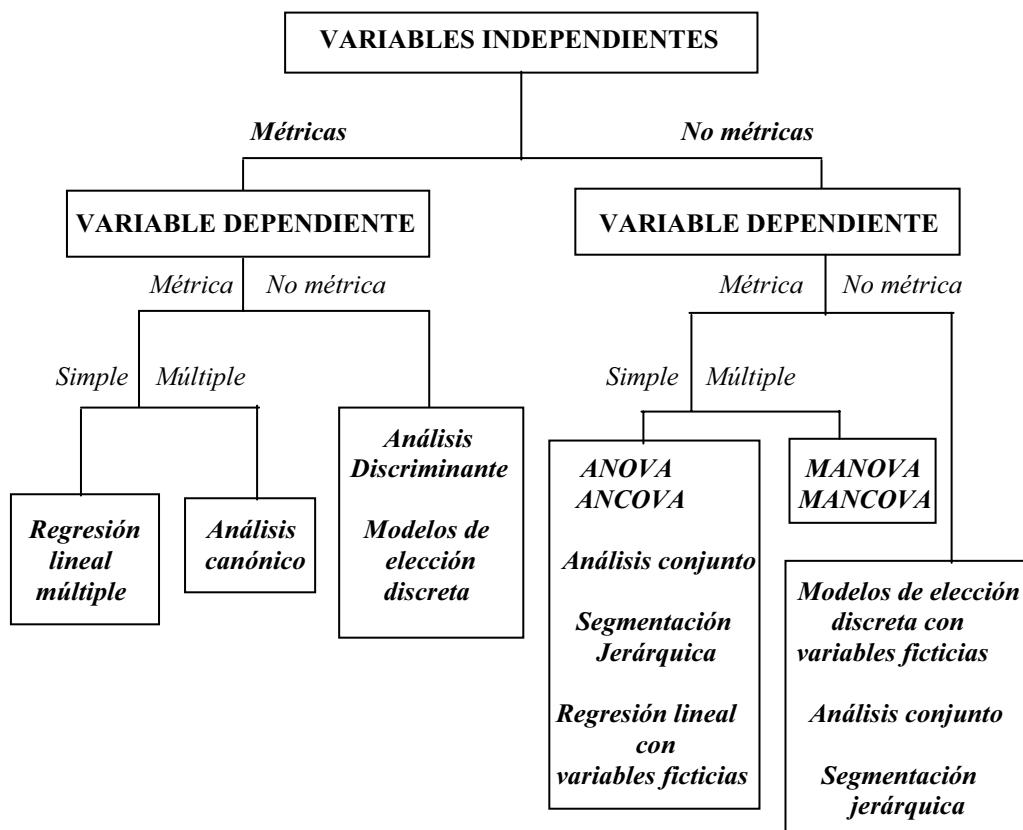
El cuadro siguiente permite clasificar las técnicas de análisis estadístico de datos simultáneamente en función del tipo de variables que manejan y del objetivo principal de su tratamiento conjunto.



MÉTODOS EXPLICATIVOS: TÉCNICAS DEL ANÁLISIS DE LA DEPENDENCIA

La clasificación de las técnicas de análisis estadístico multivariante de datos presentada en el apartado anterior comenzaba discriminando entre la existencia o no de variables explicativas y explicadas. La parte de la derecha del árbol de clasificación anterior se desarrollaba suponiendo que existía una dependencia entre las variables explicadas y sus correspondientes variables explicativas, dando lugar a los denominados *métodos explicativos*.

Con la intención de clarificar un poco más ese tipo de técnicas de análisis de la dependencia se presenta el cuadro siguiente, que las clasifica en función de la naturaleza métrica o no métrica de las variables independientes y dependientes.



Regresión múltiple

El análisis de la regresión múltiple es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) también métricas. El objetivo esencial del análisis de la regresión múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador.

La expresión funcional del análisis de la regresión múltiple es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

La regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas.

Análisis canónico (correlación canónica)

El análisis de la correlación canónica es una técnica estadística utilizada para analizar la relación entre múltiples variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) también métricas. El objetivo esencial del análisis de la correlación canónica es utilizar las variables independientes, cuyos valores son conocidos, para predecir las variables criterio (dependientes) seleccionadas por el investigador.

La expresión funcional del análisis de la correlación canónica es la siguiente:

$$G(y_1, y_2, \dots, y_n) = F(x_1, x_2, \dots, x_n)$$

El análisis de la correlación canónica también puede extenderse al caso de variables dependientes no métricas y al caso de variables independientes no métricas.

Análisis discriminante

El análisis discriminante es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas. El objetivo esencial del análisis discriminante es utilizar los valores conocidos de las variables independientes para predecir con qué categoría de la variable dependiente se corresponden. Así podremos predecir en qué categoría de riesgo crediticio se encuentra una persona, el éxito de un producto en el mercado, etc.

La expresión funcional del análisis discriminante es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

La variable dependiente y es no métrica y las variables independientes son métricas.

Formalmente podríamos decir que el análisis discriminante es una técnica de clasificación que permite agrupar a los elementos de una muestra en dos o más categorías diferentes, predefinidas en una variable dependiente no métrica, en función de una serie de variables independientes métricas combinadas linealmente.

Modelos de elección discreta

El análisis discriminante es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas, de modo que para valores conocidos de las variables independientes se predice con qué categoría (clase) de la variable dependiente se corresponden. Es decir, para valores dados de las variables independientes hemos de predecir la probabilidad de pertenencia a una categoría o clase de la variable dependiente (por ejemplo, probabilidad de que un individuo compre un producto o devuelva un crédito según algunas variables medidas en él). Los modelos de elección discreta tienen la misma naturaleza que el modelo discriminante, pero ahora lo que se predice es la probabilidad de pertenencia a una categoría (clase) para valores dados de las variables dependientes. Por lo tanto, los modelos de elección discreta predicen directamente la probabilidad de ocurrencia de un suceso que viene definido por los valores de las variables independientes.

Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que cumple esta condición se denomina **modelo lineal de probabilidad**, y tiene la forma funcional:

$$P_i = F(x_i, \beta) + u_i$$

Se observa que si F es la función de distribución de una variable aleatoria entonces P varía entre cero y uno.

En el caso particular en que la función F es la función logística estaremos ante el **modelo Logit**, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} + u_i$$

En el caso particular en que la función F es la función de distribución de una normal unitaria estaremos ante el **modelo Probit**, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\beta} e^{-\frac{t^2}{2}} dt + u_i$$

Modelo ANOVA (Análisis de la varianza simple)

El análisis de la varianza simple es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza es determinar si diversas muestras proceden de poblaciones con igual media. Los valores no métricos de las variables independientes determinarán una serie de grupos en la variable dependiente. De modo que el modelo ANOVA mide la significación estadística de las diferencias entre las medias de los grupos determinados en la variable dependiente por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza simple ANOVA es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

La variable dependiente y es métrica y las variables independientes son no métricas.

Modelo ANCOVA (Análisis de la covarianza simple)

El análisis de la covarianza simple es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas), parte de las cuales son no métricas, siendo la otra parte métricas (*covariables*).

La expresión funcional del modelo del análisis de la covarianza simple ANCOVA es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

La variable dependiente y es métrica y las variables independientes son algunas métricas y otras no métricas.

Modelo MANOVA (Análisis de la varianza múltiple)

El análisis de la covarianza simple es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medias de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza múltiple MANOVA es la siguiente:

$$G(y_1, y_2, \dots, y_m) = F(x_1, x_2, \dots, x_n)$$

Las variables dependientes son métricas y las variables independientes son no métricas.

Modelo MANCOVA (Análisis de la covarianza múltiple)

El análisis de la covarianza múltiple es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) mezcla de variables métricas y no métricas.

La expresión funcional del modelo del análisis de la covarianza múltiple MANCOVA es la siguiente:

$$G(y_1, y_2, \dots, y_m) = F(x_1, x_2, \dots, x_n)$$

Las variables dependientes son métricas y las variables independientes son una parte métricas y otra parte no métricas.

En el análisis de la covarianza, tanto simple como múltiple, las variables métricas independientes (*covariates*) tienen como objetivo eliminar determinados efectos que puedan sesgar los resultados incrementando la varianza dentro de los grupos. En el análisis de la covarianza se suele comenzar eliminando, mediante una regresión lineal, la variación experimentada por las variables dependientes producida por la covariable o covariables de efectos indeseados, para continuar con un análisis ANOVA o MANOVA sobre las variables dependientes ajustadas (residuos de la regresión anterior).

Análisis conjunto

El análisis conjunto es una técnica estadística utilizada para analizar la relación lineal o no lineal entre una variable dependiente (o endógena) generalmente ordinal (aunque también puede ser métrica) y varias variables independientes (o exógenas) no métricas.

La expresión funcional del análisis conjunto es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

La variable dependiente recoge la preferencia (intención de compra, etc.) que el individuo exhibe hacia el producto (es decir, la utilidad global que el producto le aporta) y las variables dependientes son los atributos distintivos del producto.

Es importante tener presente que sólo la variable dependiente recogerá información aportada por los individuos encuestados, ya que la información contenida en las variables independientes será especificada por el investigador en virtud de los productos que desee someter a evaluación por los encuestados.

El análisis conjunto permite generar un modelo individualizado por encuestado, de modo que el modelo general para toda la muestra resulte de la agregación de los modelos de todos los individuos que la componen. El análisis conjunto descompone las preferencias que el individuo manifiesta hacia el producto a fin de conocer qué valor le asigna a cada atributo (*técnica descomposicional*), mientras que en el análisis discriminante y en el análisis de la regresión las valoraciones de cada atributo que hace el sujeto se utilizan para componer su preferencia sobre el producto (*técnicas composicionales*).

Segmentación jerárquica

La segmentación jerárquica es una técnica que persigue distinguir grupos de elementos homogéneos en una población utilizando una variable dependiente no métrica o métrica (variable criterio para la formación de grupos) y varias variables independientes métricas que actúan como predictoras

La expresión funcional del modelo de segmentación jerárquica es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

La segmentación jerárquica también se conoce como técnica de árboles de decisión, ya que puede presentarse como un proceso iterativo descendente de partición de la muestra total en sucesivos grupos en virtud del valor adoptado por la variable dependiente, el cual es función de los valores de las variables explicativas.

Regresión múltiple y modelos de elección discreta con variables ficticias

La regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas. A cada clase de la variable no métrica se le asigna un valor numérico.

El modelo de regresión múltiple con variables ficticias es similar al análisis de la regresión múltiple con la diferencia de que las variables independientes pueden ser también no métricas. Por lo tanto, se trata de una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) métricas, no métricas o mezcla de ambas. El objetivo esencial del análisis de la regresión múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador.

La expresión funcional del análisis de la regresión múltiple con variables ficticias es la siguiente:

$$y = F(x_1, x_2, \dots, x_n)$$

Al igual que la regresión múltiple, los modelos de elección discreta admiten la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas.

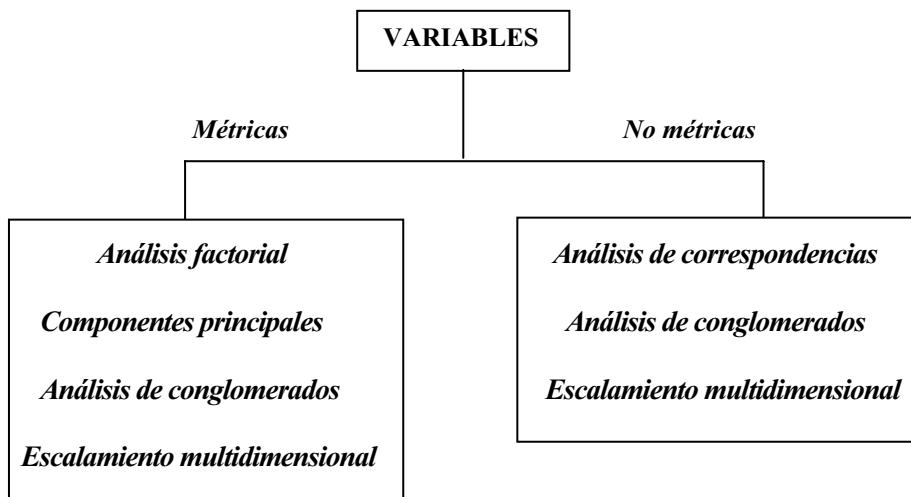
Podríamos tabular los *métodos del análisis multivariante de la dependencia, según la naturaleza de sus variables dependientes e independientes*, como sigue:

TÉCNICA	Variables dependientes	Variables independientes
ANOVA y MANOVA	Métrica (métricas)	No métricas
ANCOVA y MANCOVA	Métrica (métricas)	Métricas y no métricas
REGRESIÓN MÚLTIPLE	Métrica	Métricas
REGRESIÓN MÚLTIPLE (VARIABLES FICTICIAS)	Métrica	Métricas y no métricas
CORRELACIÓN CANÓNICA	Métricas y no métricas	Métricas y no métricas
ELECCIÓN DISCRETA	No métrica	Métricas
ELECCIÓN DISCRETA (VARIABLES FICTICIAS)	No métrica	Métricas y no métricas
ANÁLISIS CONJUNTO	Métrica o no métrica	No métricas
SEGMENTACIÓN JERÁRQUICA	Métrica o no métrica	No métricas

MÉTODOS DESCRIPTIVOS: TÉCNICAS DEL ANÁLISIS DE LA INTERDEPENDENCIA

La clasificación global de las técnicas de análisis multivariante discriminaba entre la existencia o no de variables explicativas y explicadas. La parte de la izquierda del árbol de clasificación global presentado anteriormente se desarrollaba suponiendo que no existía una dependencia entre las variables explicadas y las variables explicativas, dando lugar a los denominados *métodos descriptivos*.

Con la intención de clarificar un poco más ese tipo de técnicas de análisis de la interdependencia se presenta el cuadro siguiente, que las clasifica en función de la naturaleza métrica o no métrica de las variables.



Análisis en componentes principales

El análisis en componentes principales es una técnica multivariante que persigue reducir la dimensión de una tabla de datos excesivamente grande por el elevado número de variables que contiene x_1, x_2, \dots, x_n y quedarse con unas cuantas variables C_1, C_2, \dots, C_p combinación de las iniciales (*componentes principales*) **perfectamente calculables** y que sinteticen la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables:

$$C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)$$

$$\vdots$$

$$C_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n)$$

Pero sólo se retienen las p componentes (componentes principales) que explican un porcentaje alto de la variabilidad de las variables iniciales (C_1, C_2, \dots, C_p).

Será necesario tener en cuenta el tipo de variables que se maneja. En el análisis en componentes principales las variables tienen que ser cuantitativas. Las componentes deben de ser suficientes para resumir la mayor parte de la información contenida en las variables originales.

Asimismo cada variable original podrá expresarse en función de las componentes principales, de modo que la varianza de cada variable original se explica completamente por las componentes cuya combinación lineal la determinan.

$$\begin{aligned}x_1 &= r_{11}C_1 + r_{12}x_2 + \cdots + r_{1p}C_p \\&\vdots \\r_{ij} &= \sqrt{\lambda_i}a_{ij} \\x_n &= r_{n1}C_1 + r_{n2}C_2 + \cdots + r_{np}C_p\end{aligned}$$

Se demuestra que r_{ij} es el coeficiente de correlación entre la componente C_i y la variable x_j y se calcula multiplicando el peso a_{ij} de la variable en esa componente por la raíz cuadrada de su valor propio λ_i (cada componente principal C_i se asocia con el valor propio i -ésimo (en magnitud) de la matriz (a_{ij}) .

Análisis factorial

El análisis factorial, al igual que el análisis en componentes principales, es una técnica multivariante que persigue *reducir la dimensión de una tabla de datos excesivamente grande* por el elevado número de variables que contiene y quedarse con unas cuantas **variables ficticias** que, aunque **no observadas**, sean combinación de las reales y sinteticen la mayor parte de la información contenida en sus datos.

Aquí también será necesario tener en cuenta el tipo de variables que se maneja. En el análisis factorial las variables tienen que ser cuantitativas. Los factores deben de ser suficientes para resumir la mayor parte de la información contenida en las variables originales.

La diferencia entre análisis en componentes principales y análisis factorial radica en que en el análisis factorial se trata de encontrar variables sintéticas latentes, inobservables y aún no medidas cuya existencia se sospecha en las variables originales y que permanecen a la espera de ser halladas, mientras que en el análisis en componentes principales se obtienen variables sintéticas combinación de las originales y cuyo cálculo es posible basándose en aspectos matemáticos independientes de su interpretabilidad práctica.

En el análisis en componentes principales la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan, sus componentes. Pero esto no ocurre en el análisis factorial.

En el análisis factorial sólo una parte de la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (*factores comunes* F_1, F_2, \dots, F_p). Esta parte de la variabilidad de cada variable original explicada por los factores comunes se denomina *comunalidad*, mientras que la parte de varianza no explicada por los factores comunes se denomina *unicidad* (*comunalidad + unicidad = 1*) y representa la parte de variabilidad propia f_i de cada variable x_i .

$$\begin{aligned}x_1 &= r_{11}F_1 + r_{12}x_2 + \cdots + r_{1p}F_p + f_1 \\&\vdots \\x_n &= r_{n1}F_1 + r_{n2}F_2 + \cdots + r_{np}F_p + f_n\end{aligned}$$

Cuando la communalidad es unitaria (unicidad nula) el análisis en componentes principales coincide con el factorial. Es decir, el análisis en componentes principales es un caso particular del análisis factorial en el que los factores comunes explican el 100% de la varianza total.

Análisis de correspondencias

El análisis factorial, al igual que el análisis en componentes principales, es una técnica multivariante que persigue *reducir la dimensión de una tabla de datos* formada por *variables cuantitativas*. Si las variables fuesen *variables cualitativas*, estaríamos ante el análisis de correspondencias.

Cuando se estudia conjuntamente el comportamiento de dos variables cualitativas estamos ante el *Análisis de correspondencias simples*, pero este análisis puede ser generalizado para el caso en que se dispone de un número de variables cualitativas mayor que dos, en cuyo caso estamos ante el *Análisis de correspondencias múltiples*. En el caso de correspondencias simples los datos de las dos variables cualitativas pueden representarse en una tabla de doble entrada, denominada *tabla de contingencia*. En el caso de las correspondencias múltiples la tabla de contingencia de doble entrada pasa a ser una hipertabla en tres o más dimensiones, difícil de representar y que suele sintetizarse en la denominada *tabla de Burt*.

El objetivo del análisis de correspondencias es establecer relaciones entre variables no métricas enriqueciendo la información que ofrecen las tablas de contingencia, que sólo comprueban si existe alguna relación entre las variables (test de la chi-cuadrado, etc.) y la intensidad de dicha relación (test V de Cramer, etc.). El análisis de correspondencias revela además en qué grado contribuyen a esa relación detectada los distintos valores de las variables, información que suele ser proporcionada en modo gráfico (valores asociados próximos).

Podríamos sintetizar diciendo que el análisis de correspondencias busca como objetivo el estudio de la asociación entre las categorías de múltiples variables no métricas, pudiendo obtenerse un mapa perceptual que ponga de manifiesto esta asociación en modo gráfico.

Análisis de conglomerados (análisis cluster)

El análisis de conglomerados es una técnica estadística multivariante de clasificación automática de datos, que a partir de una tabla de casos-variables, trata de situar todos los casos en grupos homogéneos (conglomerados o clusters) no conocidos de antemano pero sugeridos por la propia esencia de los datos, de manera que individuos que puedan ser considerados similares sean asignados a un mismo cluster, mientras que individuos diferentes (disimilares) se sitúen en clusters distintos.

La creación de grupos basados en similaridad de casos exige una definición de similaridad o de su complementario (distancia entre individuos). Existen muchas formas de medir estas distancias y diferentes reglas matemáticas para asignar los individuos a distintos grupos, dependiendo del fenómeno estudiado y del conocimiento previo de posible agrupamiento que se tenga.

El análisis de conglomerados suele comenzar estimando las similitudes entre los individuos (u objetos) a través de correlación (distancia o asociación) de las distintas variables (métricas o no métricas) de que se dispone. A continuación se establece un procedimiento que permite comparar los grupos en virtud de las similitudes. Por último se decide cuántos grupos se construyen, teniendo en cuenta que cuanto menor sea el número de grupos, menos homogéneos serán los elementos que integran cada grupo. Se perseguirá formar el mínimo número de grupos lo más homogéneos posibles dentro de sí y lo más heterogéneos posibles entre sí.

El análisis cluster se diferencia del análisis factorial en que en el análisis factorial se constituyen los factores agrupando variables, mientras que en el análisis cluster se constituyen los conglomerados agrupando individuos (objetos) o también variables. Al aplicar análisis factorial en un factor determinado se incluyen variables que están relacionadas con él (positiva y negativamente), pero en el análisis cluster, las variables relacionadas positivamente forman parte de un conglomerado distinto del de las variables relacionadas negativamente.

El análisis factorial, al igual que el análisis en componentes principales, es una técnica multivariante que persigue *reducir la dimensión de una tabla de datos* formada por *variables cuantitativas*. Si las variables fuesen *variables cualitativas*, estaríamos ante el análisis de correspondencias. Un análisis cluster puede complementarse con un análisis discriminante que, una vez identificados los conglomerados, verifique si existe una relación causal o no entre la pertenencia a un conglomerado determinado y los valores de las variables.

Escalamiento multidimensional

El escalamiento multidimensional tiene como finalidad crear una representación gráfica (*mapa perceptual*) que permita conocer la situación de los individuos en un conjunto de objetos por posicionamiento de cada uno en relación a los demás. Dicha situación será producto de las percepciones y preferencias o similitudes entre los objetos apreciadas por los sujetos. Estas percepciones (preferencias o similitudes) son la entrada del análisis, y pueden ser variables métricas o no métricas. El escalamiento multidimensional transforma estas variables en distancias entre los objetos en un espacio de dimensiones múltiples, de modo que objetos que aparecen situados más próximos entre sí son percibidos como más similares por los sujetos.

Existe una diferencia clave entre el escalamiento multidimensional y el análisis cluster. En el escalamiento multidimensional se desconocen los elementos de juicio de los encuestados y no se conocen las variables que implícitamente están considerando éstos para realizar su evaluación de las preferencias por los objetos. En el análisis cluster las similitudes entre objetos se obtienen a partir de una combinación de variables estudiadas.

El escalamiento multidimensional es de más fácil aplicación que el análisis factorial, ya que no requiere supuestos de linealidad, ni que las variables sean métricas, ni un tamaño mínimo de muestra.

Resumiendo, podríamos definir el escalamiento multidimensional como una técnica cuyo fin es elaborar una representación gráfica que permita conocer la imagen que los individuos se crean de un conjunto de objetos por posicionamiento de cada uno en relación a los demás (*mapa perceptual*).

Podríamos tabular los *métodos del análisis multivariante de la interdependencia, según la naturaleza de sus variables y los grupos que se forman*, como sigue:

TÉCNICA	Variables	Se forman grupos de
COMPONENTES PRINCIPALES	Métricas	Variables
ANÁLISIS FACTORIAL	Métricas	Variables
ANÁLISIS DE CORRESPONDENCIAS	No métricas	Categorías de variables
ANÁLISIS DE CONGLOMERADOS	Métricas y no métricas	Objetos
ESCALAMIENTO MULTIDIMENSIONAL	Métricas y no métricas	Objetos

TÉCNICAS EMERGENTES DE ANÁLISIS MULTIVARIANTE DE DATOS

La disponibilidad de grandes volúmenes de datos y el uso generalizado de herramientas informáticas ha transformado el análisis multivariante orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de **Minería de datos** o **Data Mining**.

El Data Mining puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.

Las técnicas de Data Mining persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas estadísticas avanzadas de análisis multivariante de datos.

Pero por otro lado podemos decir que las técnicas de Data Mining son tan antiguas como la estadística misma. De hecho, las técnicas estadísticas que utiliza el Data Mining coinciden en su mayoría con las técnicas estadísticas de análisis multivariante de datos. La clasificación inicial de las técnicas de Data Mining distingue entre técnicas de modelado originado por la teoría en las que las variables pueden clasificarse inicialmente en dependientes e independientes (similares a las técnicas del análisis de la dependencia o métodos explicativos del análisis multivariante), técnicas de modelado originado por los datos en las que todas las variables tienen inicialmente el mismo estatus (similares a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante) y técnicas auxiliares.

Las **técnicas de modelado originado por la teoría** especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de Data Mining antes de aceptarlo como válido. Formalmente, la aplicación de todo modelo debe superar las fases de **identificación objetiva** (a partir de los datos se aplican reglas que permitan identificar el mejor modelo posible que ajuste los datos), **estimación** (proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación), **diagnóstico** (proceso de contraste de la validez del modelo estimado) y **predicción** (proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes). Podemos incluir entre estas técnicas todos los tipos de regresión y asociación, análisis de la varianza y covarianza, análisis discriminante y series temporales.

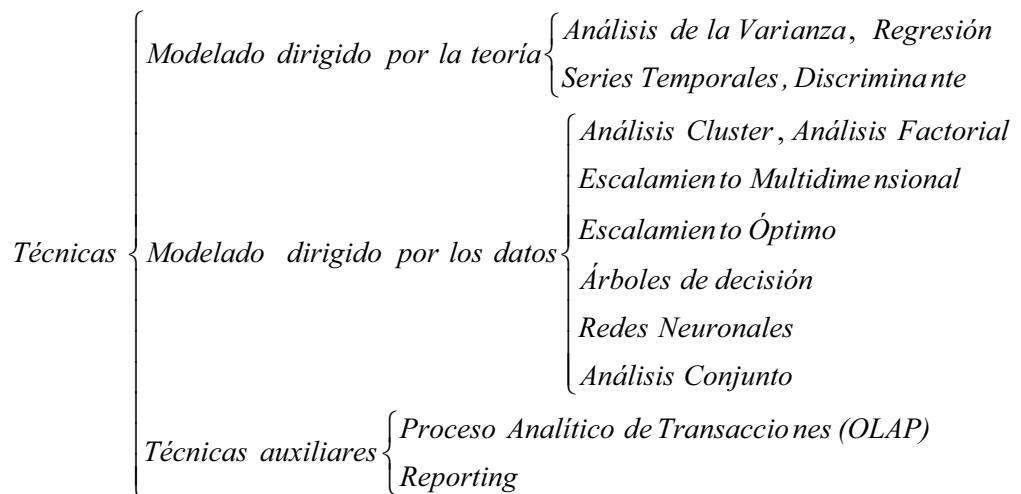
En las *técnicas de modelado originado por los datos* no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. El modelo se obtiene como mezcla del conocimiento obtenido antes y después del Data Mining y también debe contrastarse antes de aceptarse como válido. Por ejemplo, las *redes neuronales* permiten descubrir modelos complejos y afinarlos a medida que progresá la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa.

Por su parte, las *técnicas de clasificación* extraen perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato. Asimismo, los *árboles de decisión* permiten dividir datos en grupos basados en los valores de las variables. Esta técnica permite determinar las variables significativas para un elemento dado. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas.

Además de las redes neuronales, los árboles de decisión y las técnicas de clasificación (cluster, etc.), podemos incluir en este grupo las *técnicas de reducción de la dimensión* (factorial, componentes principales, correspondencias, etc.), las técnicas de escalamiento óptimo y multidimensional y el análisis conjunto.

Las *técnicas auxiliares* son herramientas más superficiales y limitadas. Se trata de nuevos métodos basados en técnicas estadísticas descriptivas e informes.

A continuación se muestra una *clasificación de las técnicas de Data Mining*.



FASES A SEGUIR EN LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE DE DATOS

Para llevar a cabo con éxito la aplicación de cualquier técnica de análisis multivariante deben resolverse asuntos que van desde el problema de definición del modelo hasta un diagnóstico crítico de los resultados. La aproximación a la modelización se centra en el análisis de un plan de investigación bien definido, comenzando por un modelo conceptual que detalle las relaciones a examinar. Definido el modelo, se pueden iniciar los trabajos empíricos, incluyendo la selección de una técnica multivariante específica y su puesta en práctica. Después de haber obtenido resultados significativos, el asunto central es la interpretación. Finalmente, las medidas de diagnosis aseguran que el modelo no sólo es válido para la muestra de datos sino que es también generalizable.

Un primer paso en la práctica del análisis multivariante es *definir el problema de investigación, objetivos y técnica multivariante conveniente*. El investigador debe ver en primer lugar el problema en términos conceptuales, definiendo los conceptos e identificando las relaciones fundamentales a investigar. Si se propone un método de dependencia el investigador debe especificar los conceptos de dependencia e independencia. Si se propone una técnica de interdependencia se deben determinar las dimensiones de la estructura o similitud. Con los objetivos y el modelo conceptual especificados, el analista sólo tiene que elegir la técnica multivariante apropiada.

Un segundo paso en la práctica del análisis multivariante es *desarrollar el proyecto de análisis poniendo en práctica la técnica multivariante*. Con el modelo conceptual establecido, el investigador debe poner en práctica la técnica seleccionada. Para cada técnica, el investigador debe desarrollar un plan de análisis específico que dirija el conjunto de supuestos que subyacen en la aplicación de la técnica. Estos supuestos pueden ser el tamaño de muestra mínimo deseado, los tipos de variables permitidas, métodos de estimación, etc. Estos supuestos resuelven los detalles específicos y finalizan la formulación del modelo y los requisitos del esfuerzo de recogida de datos.

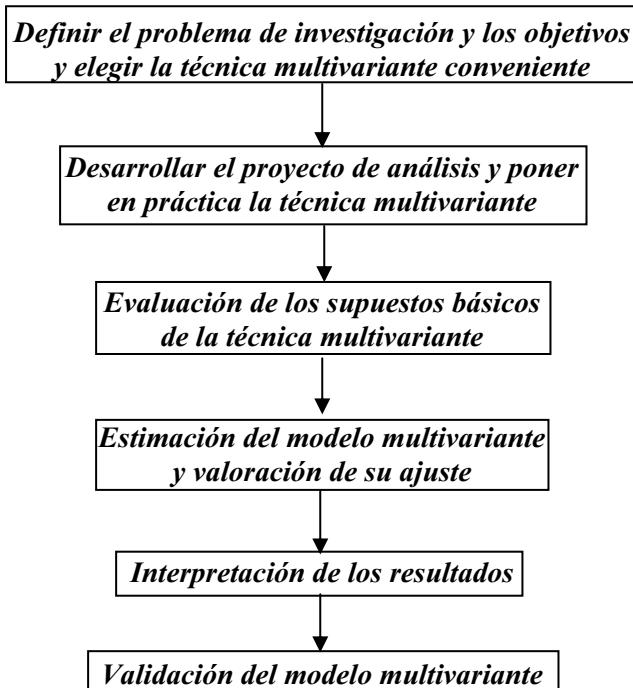
Un tercer paso en la práctica del análisis multivariante es la *evaluación de los supuestos básicos de la técnica multivariante*. Una vez recogidos los datos, el primer paso del análisis no es estimar el modelo, sino evaluar que se cumplen los supuestos subyacentes. Para las técnicas de la dependencia será necesario comprobar los supuestos de normalidad, linealidad, homoscedasticidad, etc. antes de aplicar el modelo. Para las técnicas de la interdependencia se comprobarán, entre otros, los supuestos de correlación entre las variables.

Un cuarto paso en la práctica del análisis multivariante es la **estimación del modelo multivariante y la valoración del ajuste del modelo**. Una vez satisfechos todos los supuestos del modelo, se procede a su estimación efectiva realizando a continuación una valoración global del ajuste del modelo (parámetros significativos individual y globalmente, capacidad de predicción del modelo, etc.).

Un quinto paso en la práctica del análisis multivariante es la **interpretación de los valores obtenidos**. Una vez estimado el modelo será necesario interpretar los resultados de acuerdo a los valores teóricos posibles. Esta interpretación puede reconducir a la reespecificación del modelo y a su nueva estimación hasta que la interpretación de los resultados se ajuste coherentemente a los valores teóricos. Hasta que no se cumpla esta condición, no existe evidencia empírica de que las relaciones multivariantes de los datos muestrales puedan generalizarse para toda la población.

Un sexto paso en la práctica del análisis multivariante es la **validación del modelo multivariante**. Una vez estimado el modelo será necesario aceptar los resultados con grado de fiabilidad lo más alto posible mediante la aplicación de contrastes específicos de cada técnica.

Las fases anteriores pueden esquematizarse como sigue:



PRIMEROS PASOS EN EL ANÁLISIS MULTIVARIANTE. ANÁLISIS EXPLORATORIO DE DATOS

ANÁLISIS PREVIO DE LOS DATOS

Antes de aplicar cualquier técnica de análisis multivariante es preciso realizar un análisis previo de los datos de que se dispone. Es necesario examinar las variables individuales y las relaciones entre ellas, así como evaluar y solucionar problemas en el diseño de la investigación y en la recogida de datos tales como el tratamiento de la información faltante (o datos ausentes) y la presencia de datos anómalos (o atípicos).

La primera tarea que se suele abordar es el *análisis exploratorio y gráfico de los datos*. La mayoría del software estadístico dispone de herramientas que aportan técnicas gráficas preparadas para el examen de los datos que se ven mejoradas con medidas estadísticas más detalladas para su descripción. Estas técnicas permiten el examen de las características de la distribución de las variables implicadas en el análisis, las relaciones bivariantes (y multivariantes) entre ellas y el análisis de las diferencias entre grupos. Hay que tener presente que las representaciones gráficas nunca sustituyen a las medidas de diagnóstico formal estadístico (contrastos de ajuste de los datos a una distribución, contrastes de asimetría, contrastes de aleatoriedad, etc.), pero proporcionan una forma alternativa de desarrollar una perspectiva del carácter de los datos y de las interrelaciones que existen, incluso si son multivariantes.

La segunda tarea que suele llevarse a cabo antes de aplicar cualquier técnica multivariante es el *análisis de los datos ausentes*. Cualquier recogida y proceso de datos presenta problemas que van a impedir obtener información de algunos de los elementos de la población en estudio.

Entre estos problemas cabría destacar las negativas a colaborar, las ausencias de los encuestados en el momento de la toma de datos, la inaccesibilidad a algunos elementos o los errores en los instrumentos de medida. La existencia de datos ausentes nunca debe impedir la aplicación del análisis multivariante. Tampoco debe limitar la posibilidad de generalizar los resultados de una investigación. El analista deberá identificar la presencia de datos ausentes y llevar a cabo las acciones necesarias para intentar minimizar sus efectos.

La tercera tarea para aplicar cualquier técnica multivariante es la **detección de valores atípicos**. Se trata de detectar la existencia de observaciones que no siguen el mismo comportamiento que el resto. Los casos atípicos suelen deberse a errores en el procedimiento a la hora de introducir los datos o de codificarlos. También pueden ser consecuencia de acontecimientos anormales que hacen destacar determinadas observaciones. Una vez detectados los casos atípicos el analista debe saber elegir con criterio entre eliminarlos del análisis o evaluar toda la información incluyéndolos. Los valores atípicos también suelen denominarse *outliers*.

Una última tarea previa a la aplicación de las técnicas multivariantes es la **comprobación de los supuestos subyacentes en los métodos multivariantes**. Estos supuestos suelen ser el contraste de la *normalidad* de todas y cada una de las variables que forman parte del estudio, el testeo de la *linealidad* de las relaciones entre las variables que intervienen en el estudio (la relación entre la posible variable dependiente y las variables independientes que la explican ha de ser una ecuación lineal), la comprobación de la *homoscedasticidad* de los datos que consiste en ver que la variación de la variable dependiente que se intenta explicar a través de las variables independientes no se concentra en un pequeño grupo de valores independientes (se tratará por tanto de ver la igualdad de varianzas para los datos agrupados según valores similares de la variable dependiente) y la comprobación de la *multicolinealidad* o existencia de relaciones entre las variables independientes. A veces también es necesario contrastar la ausencia de *correlación serial de los residuos o autocorrelación*, que consiste en asegurar que cualquiera de los errores de predicción no está correlacionado con el resto

ANÁLISIS EXPLORATORIO Y GRÁFICO DE LOS DATOS

Actualmente se utilizan las novedosas técnicas del análisis exploratorio de datos, mediante las cuales se pueden analizar los datos exhaustivamente y detectar las posibles anomalías que presentan las observaciones. J. W. Tukey ha sido uno de los pioneros en la introducción de este tipo de análisis. Los estadísticos descriptivos más habitualmente utilizados han sido la media y la desviación típica. Sin embargo, el uso automático de estos índices no es muy aconsejable. La media y la desviación típica son índices convenientes sólo cuando la distribución de datos es aproximadamente normal o, al menos, simétrica y unimodal. Pero las variables objeto de estudio no siempre cumplen estos requisitos. Por lo tanto es necesario un examen a fondo de la estructura de los datos.

Se recomienda iniciar un análisis exploratorio de datos con gráficos que permitan visualizar su estructura. Por ejemplo, para datos cuantitativos es aconsejable comenzar con el gráfico de tallo y hojas o histograma digital. El paso siguiente suele ser examinar la posible presencia de normalidad, simetría y valores atípicos (*outliers*) en el conjunto de datos. Para ello suelen utilizarse los gráficos de caja y bigote. No obstante los gráficos de caja siempre deben ir acompañados de los histogramas digitales (o gráficos de tallo y hojas), ya que los primeros no detectan la presencia de distribuciones multimodales. Los gráficos de dispersión nos dan una idea de las relaciones entre variables y su ajuste.

El uso de *estadísticos robustos* (o resistentes) es muy aconsejable cuando los datos no se ajustan a una distribución normal. Estos estadísticos son los que se ven poco afectados por valores atípicos. Suelen estar basados en la mediana y en los cuartiles y son de fácil cálculo. Fruto del análisis exploratorio, a veces es necesario *transformar las variables*.

Histograma de frecuencias

De todas formas, siempre es conveniente iniciar el análisis exploratorio de datos con la construcción del histograma de frecuencias asociado, para poder así intuir la distribución de probabilidad de los datos, su normalidad, su simetría y otras propiedades interesantes en el análisis de datos.

Como ejemplo podemos considerar la variable X definida como el consumo de combustible en litros a los 1 000 kilómetros de los automóviles de una determinada marca. Los valores para X son los siguientes:

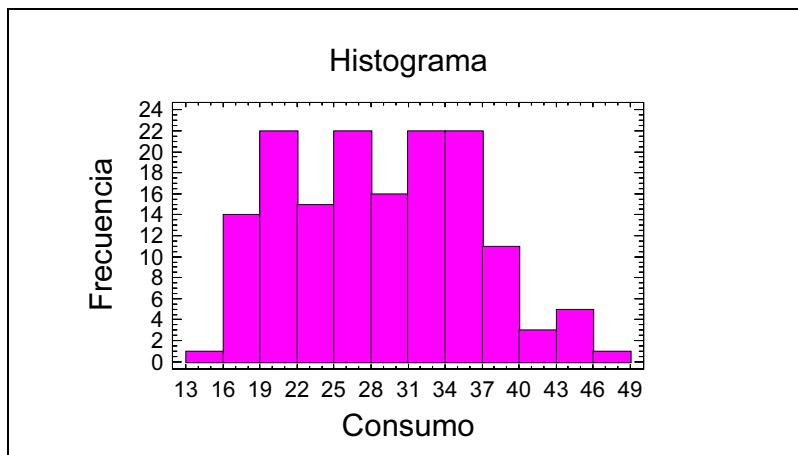
43,1	36,1	32,8	39,4	36,1	19,9	19,4	20,2	19,2	20,5	20,2	25,1	20,5	19,4	20,6
20,8	18,6	18,1	19,2	17,7	18,1	17,5	30	27,5	27,2	30,9	21,1	23,2	23,8	23,9
20,3	17	21,6	16,2	31,5	29,5	21,5	19,8	22,3	20,2	20,6	17	17,6	16,5	18,2
16,9	15,5	19,2	18,5	31,9	34,1	35,7	27,4	25,4	23	27,2	23,9	34,2	34,5	31,8
37,3	28,4	28,8	26,8	33,5	41,5	38,1	32,1	37,2	28	26,4	24,3	19,1	34,3	29,8
31,3	37	32,2	46,6	27,9	40,8	44,3	43,4	36,4	30,4	44,6	40,9	33,8	29,8	32,7
23,7	35	23,6	32,4	27,2	26,6	25,8	23,5	30	39,1	39	35,1	32,3	37	37,7
34,1	34,7	34,4	29,9	33	34,5	33,7	32,4	32,9	31,6	28,1	30,7	25,4	24,2	22,4
26,6	20,2	17,6	28	27	34	31	29	27	24	23	36	37	31	38
36	36	36	34	38	32	38	25	38	26	22	32	36	27	27
44	32	28	31											

Para explorar esta información elaboramos la tabla de frecuencias asociada a los datos y estudiamos la posible normalidad y simetría de la distribución del consumo de combustible. Observamos que tenemos 154 valores sobre el consumo de los automóviles que inicialmente no aportan mucha información. Evidentemente hay una variabilidad en el consumo de los automóviles; sin embargo, es muy difícil detectar qué patrón sigue dicha variabilidad para determinar mejor la estructura de los datos. En primer lugar será conveniente realizar una ordenación de los datos según su magnitud, es decir, una tabla de frecuencias, que aportará algo de luz sobre la distribución de frecuencias subyacente.

Como se trata de una variable cuantitativa con 154 valores comprendidos entre 13 y 49, será necesario agruparlos en intervalos o clases. Para ello tomamos 12 intervalos de igual anchura (12 es un entero que aproxima bien la raíz cuadrada de $N = 154$). La anchura de los intervalos será $(49 - 13)/12 = 3$. Se obtiene la siguiente tabla de frecuencias:

Intervalo	Límite inferior	Límite superior	Marca de clase	n_i	$f_i = n_i/N$	N_i	$F_i = n_i/N$
1	13,0	16,0	14,5	1	0,0065	1	0,0065
2	16,0	19,0	17,5	14	0,0909	15	0,0974
3	19,0	22,0	20,5	22	0,1429	37	0,2403
4	22,0	25,0	23,5	15	0,0974	52	0,3377
5	25,0	28,0	26,5	22	0,1429	74	0,4805
6	28,0	31,0	29,5	16	0,1039	90	0,5844
7	31,0	34,0	32,5	22	0,1429	112	0,7273
8	34,0	37,0	35,5	22	0,1429	134	0,8701
9	37,0	40,0	38,5	11	0,0714	145	0,9416
10	40,0	43,0	41,5	3	0,0195	148	0,9610
11	43,0	46,0	44,5	5	0,0325	153	0,9935
12	46,0	49,0	47,5	1	0,0065	154	1,0000

La siguiente tarea es la construcción del histograma de frecuencias, gráfico adecuado para una variable cuantitativa con sus valores agrupados en intervalos. Su representación es la siguiente:



Se observa que la distribución subyacente que modela los datos sobre la variable consumo de los automóviles es aproximadamente simétrica y ajustable a forma de campana, lo que permite pensar en la existencia de normalidad y simetría en la distribución de X.

Vemos así que el histograma da una idea clara de la distribución de la variable, incluyendo un modelo probabilístico para su modelación, en este caso la distribución normal. El simple examen de los datos tabulados inicialmente no aportaba información alguna, sin embargo su graficación da luz al proceso.

Diagrama de tallo y hojas

El diagrama de tallo y hojas es un procedimiento semigráfico para presentar la información para variables cuantitativas, que es especialmente útil cuando el número total de datos es pequeño (menor que 50). Los principios para la realización del diagrama (debido a Tukey) son los siguientes:

- Redondear los datos a dos o tres cifras significativas.
- Disponerlos en dos columnas separadas por una línea vertical de tal forma que para los datos con dos dígitos la cifra de las decenas se encuentre a la izquierda de la línea vertical (tallo del diagrama), y a la derecha las unidades (hojas o ramas del diagrama). Por ejemplo, 87 se escribirá 8 7. Para datos con tres dígitos el tallo estará formado por los dígitos de las centenas y las decenas, que se escribirán a la izquierda de la línea vertical, y las hojas estarán formadas por el dígito de las unidades, que se escribirá a la derecha de la línea vertical.
- Cada tallo define una clase, y se escribe sólo una vez. A su derecha se van escribiendo por orden las sucesivas hojas correspondientes a ese tallo. El número de hojas para cada tallo representa la frecuencia de cada clase.

El diagrama de tallo y hojas, también llamado *histograma digital*, es una combinación entre un histograma de barras y una tabla de frecuencias. Al mantener los valores de la variable, el diagrama de tallo y hojas resulta más informativo que el clásico histograma de barras, ya que conserva los datos originales y, al mismo tiempo, compone un perfil que ayuda a estudiar la forma y simetría de la distribución. Se trata pues de una herramienta de análisis exploratorio de datos que muestra el rango de los datos, dónde están más concentrados, su simetría y la presencia de datos atípicos. Este procedimiento no es muy aconsejable para conjuntos de datos grandes.

A continuación se presenta el diagrama de tallo y hojas para la variable X relativa a la variable consumo de los automóviles definida en el apartado anterior.

Diagrama de Tallo y Hojas para X: unidad = 1,0 1|2 representa 12,0

23	1 56667777788889999999
51	2 0000000001112223333333444
(31)	2 555566666777777788889999
72	3 0000011111111222222222333344444444
34	3 55566666666777778888999
9	4 00133444
1	4 6

El rango de X ha sido dividido en 7 clases o intervalos llamados *tallos*, cada uno de ellos representado por una fila del diagrama. El primer número de cada fila (separado de los demás) presenta la frecuencia absoluta de la clase correspondiente. El segundo número de cada fila presenta la cifra de las decenas de cada valor de X en su correspondiente clase. El resto de los números de cada fila (llamados *hojas*) son las cifras de las unidades de todos los elementos de la clase definida por la fila. De esta forma, además de presentar la distribución de los elementos en forma de histograma horizontal, en el diagrama se observan los propios elementos. Las hojas permiten analizar la simetría, la normalidad y otras características de la distribución de igual forma que un histograma.

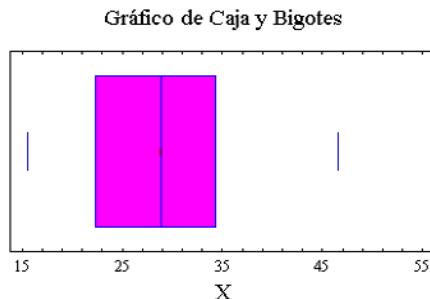
Gráfico de caja y bigotes

El gráfico de caja y bigotes permite analizar y resumir un conjunto de datos univariante dado. Esta herramienta de análisis exploratorio de datos va a permitir estudiar la simetría de los datos, detectar valores atípicos y vislumbrar un ajuste de los datos a una distribución de frecuencias determinada.

El gráfico de caja y bigotes divide los datos en cuatro áreas de igual frecuencia, una caja central dividida en dos áreas por una línea vertical y otras dos áreas representadas por dos segmentos horizontales (bigotes) que parten del centro de cada lado vertical de la caja. La caja central encierra el 50 por ciento de los datos. El sistema dibuja la mediana como una línea vertical en el interior de la caja. Si esta línea está en el centro de la caja no hay asimetría en la variable. Los lados verticales de la caja están situados en los cuartiles inferior y superior de la variable. Partiendo del centro de cada lado vertical de la caja se dibujan los dos bigotes, uno hacia la izquierda y el otro hacia la derecha. El bigote de la izquierda tiene un extremo en el primer cuartil Q_1 , y el otro en el valor dado por el primer cuartil menos 0,5 veces el rango intercuartílico, esto es, $Q_1 - 1,5 * (Q_3 - Q_1)$.

El bigote de la derecha tiene un extremo en el tercer cuartil Q_3 y el otro en el valor dado por el tercer cuartil más 1,5 veces el rango intercuartílico, esto es, $Q_3 + 1,5 * (Q_3 - Q_1)$. El sistema considera valores atípicos (*outliers*) los que se encuentren a la izquierda del bigote izquierdo y a la derecha del bigote derecho. El sistema separa estos datos del resto y los representa mediante puntos alineados con la línea horizontal central para que sean fáciles de detectar. En el interior de la caja central se representa la media con un signo más.

A continuación se presenta el Gráfico de caja y bigotes para la variable X relativa a la variable consumo de los automóviles.



El gráfico permite afirmar que la variable X (consumo de los automóviles cada 1000 kilómetros) varía entre 15,5 y 46,6 y que el 50% central de los coches consume entre 22 (primer cuartil) y 34,5 (tercer cuartil) litros a los 1000 kilómetros. Por otra parte, no existen valores de X anormalmente grandes (*outliers*), ya que en la Figura no aparecen puntos alineados con los bigotes. La distribución es ligeramente asimétrica hacia la derecha, ya que la zona de la derecha en el área central de la Figura es mayor que la de la izquierda. La mediana corresponde aproximadamente al valor 29 de X.

Los gráficos de caja y bigotes permiten la opción de representación *Muesca de mediana*, que sitúa dos muescas sobre los extremos de la mediana.

La anchura de la muesca representa un intervalo de confianza aproximado para la mediana con un coeficiente de confianza del 95% ($\alpha = 0,05$) que viene determinado por la expresión $M \pm (1,25R/1,35\sqrt{n}) (1+1/\sqrt{2})Z_{\alpha/2}/2$, donde R es el rango intercuartílico de la variable, M es la mediana y $Z_{\alpha/2}$ es el valor de la distribución normal (0,1) que deja a la derecha $\alpha/2$ de probabilidad.

Para nuestra variable X, consumo de los automóviles, tendremos el siguiente gráfico de caja y bigotes con muescas:

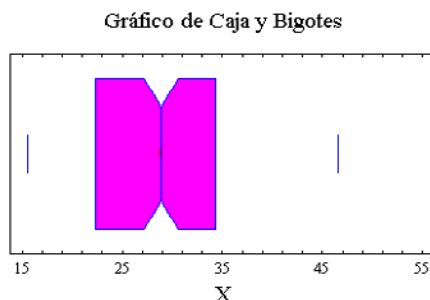


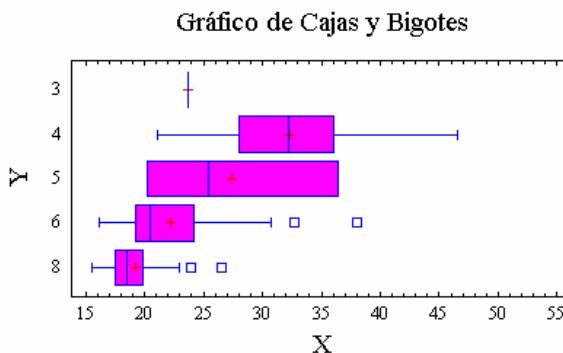
Gráfico múltiple de caja y bigotes

En estadística es típico dividir el conjunto de datos de una variable en subgrupos racionales, que pueden ser por ejemplo estratos definidos según una determinada variable de estratificación. El gráfico múltiple de caja y bigotes va a permitir analizar, resumir y comparar simultáneamente varios conjuntos de datos univariantes dados, correspondientes a los diferentes grupos en que se pueden subdividir los valores de una variable. Esta herramienta de análisis exploratorio de datos va a permitir estudiar la simetría de los datos, detectar valores atípicos y representar medias, medianas, rangos y valores extremos para todos los grupos. Al ser la representación simultánea para todos los conjuntos de datos, se podrán comparar medias, medianas, rangos, valores extremos, simetrías y valores atípicos de todos los grupos. El gráfico múltiple representará horizontalmente un gráfico de caja y bigotes para cada grupo de valores de la variable en estudio.

Si clasificamos el consumo de los automóviles (variable X) según su cilindrada (variable Y), estamos haciendo subgrupos con los valores de X según la variable de estratificación Y. Los posibles valores de Y son 8, 6, 5, 4 y 3 cilindros. Los valores de X para cada valor de Y vienen dados a continuación:

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
43,1	4	36,1	4	32,8	4	39,4	4	36,1	4	19,9	8	19,4	8	20,2	8	19,2	6	20,5	6
20,2	6	25,1	4	20,5	6	19,4	6	20,6	6	20,8	6	18,6	6	18,1	6	19,2	8	17,7	6
18,1	8	17,5	8	30	4	27,5	4	27,2	4	30,9	4	21,1	4	23,2	4	23,8	4	23,9	4
20,3	5	17	6	21,6	4	16,2	6	31,5	4	29,5	4	21,5	6	19,8	6	22,3	4	20,2	6
20,6	6	17	8	17,6	8	16,5	8	18,2	8	16,9	8	15,5	8	19,2	8	18,5	8	31,9	4
34,1	4	35,7	4	27,4	4	25,4	5	23	8	27,2	4	23,9	8	34,2	4	34,5	4	31,8	4
37,3	4	28,4	4	28,8	6	26,8	6	33,5	4	41,5	4	38,1	4	32,1	4	37,2	4	28	4
26,4	4	24,3	4	19,1	6	34,3	4	29,8	4	31,3	4	37	4	32,2	4	46,6	4	27,9	4
40,8	4	44,3	4	43,4	4	36,4	5	30,4	4	44,6	4	40,9	4	33,8	4	29,8	4	32,7	6
23,7	3	35	4	23,6	4	32,4	4	27,2	4	26,6	4	25,8	4	23,5	6	30	4	39,1	4
39	4	35,1	4	32,3	4	37	4	37,7	4	34,1	4	34,7	4	34,4	4	29,9	4	33	4
34,5	4	33,7	4	32,4	4	32,9	4	31,6	4	28,1	4	30,7	6	25,4	6	24,2	6	22,4	6
26,6	8	20,2	6	17,6	6	28	4	27	4	34	4	31	4	29	4	27	4	24	4
23	4	36	4	37	4	31	4	38	4	36	4	36	4	36	4	34	4	38	4
32	4	38	4	25	6	38	6	26	4	22	6	32	4	36	4	27	4	27	4
44	4	32	4	28	4	31	4												

El gráfico múltiple de caja y bigotes del consumo de los automóviles (variable X) según su cilindrada (variable Y), presenta el siguiente aspecto:



El Gráfico permite afirmar que la variable X (litros consumidos a los 1000 kilómetros) para los coches de 8 cilindros varía entre 15,5 y 23, y que el 50% central de estos coches consume entre 17,5 (primer cuartil) y 20 (tercer cuartil) litros a los 1 000 kilómetros, existiendo 2 valores de X anormalmente grandes (*outliers*), ya que en la Figura aparecen dos puntos alineados con el bigote de la parte derecha. La distribución de X para los coches de 8 cilindros es ligeramente asimétrica hacia la derecha, ya que la zona de la derecha en el área central de la Figura es mayor que la de la izquierda, y la mediana corresponde aproximadamente al valor 18,5 de X, siendo la media 19,5 aproximadamente.

Para los coches de 6 cilindros los litros consumidos cada 1 000 kilómetros (variable X) varían entre 16 y 31, concentrándose el 50% central de los valores de X entre 19 (primer cuartil) y 24 (tercer cuartil), existiendo 2 valores atípicos de X anormalmente grandes (*outliers*), ya que en la Figura aparecen dos puntos alineados con el bigote de la parte derecha. La distribución de X para los coches de 6 cilindros es asimétrica hacia la derecha, la mediana de X se aproxima a 20,5 y la media a 22,5.

Para los coches de 5 cilindros los litros consumidos cada 1 000 kilómetros (variable X) varían entre 20,2 y 36,5, concentrándose el 50% central de los valores de X entre los mismos valores, no existiendo bigotes ni *outliers*. La distribución de X para los coches de 5 cilindros es asimétrica hacia la derecha, la mediana de X se aproxima a 25,5 y la media a 27,5.

Para los coches de 4 cilindros los litros consumidos cada 1000 kilómetros (variable X) varían entre 21 y 47, concentrándose el 50% central de los valores de X entre 28 (primer cuartil) y 36 (tercer cuartil), no existiendo *outliers*. La distribución de X para los coches de 4 cilindros es prácticamente simétrica con valores de mediana y media aproximados a 32.

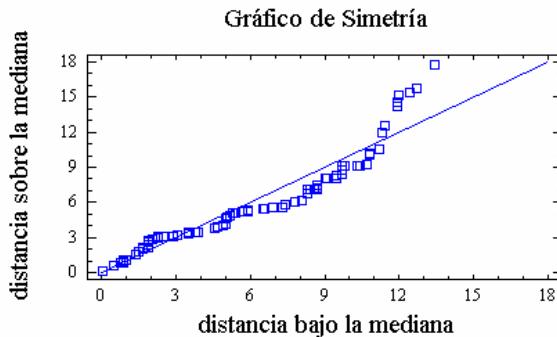
Para los coches de 3 cilindros hay un único valor de X, lo que no permite construir el gráfico de caja y bigotes.

Si comparamos los distintos gráficos, vemos que la asimetría de X es más fuerte para los coches de 5 y 6 cilindros, para los de 8 es menor y para los de 4 no existe. Valores de X anormalmente grandes sólo aparecen para los coches de 6 y 8 cilindros. Las medias y las medianas varían bastante para los diferentes grupos de valores de X determinados por los valores de Y.

Gráfico de simetría

El gráfico de simetría es una herramienta que permite analizar visualmente el grado de simetría de una variable. En el eje de abscisas se representan las distancias de los valores de la variable a la mediana que quedan por debajo de ella, y en el eje de ordenadas se representan las distancias de los valores de la variable a la mediana que quedan por encima de ella. Si la simetría fuese perfecta, el conjunto de puntos resultante sería la diagonal principal. Mientras más se aproxime la gráfica a la diagonal más simetría existirá en la distribución de la variable.

Para el ejemplo de la variable X, variable definida por el número de litros consumidos por los automóviles cada 1000 kilómetros que venimos considerando durante todo el Capítulo, tenemos el Gráfico de simetría siguiente:



Para la variable X , se observa un buen grado de simetría, ya que los puntos de la gráfica se ajustan bien a la diagonal.

Los pasos prácticos para elaborar el gráfico de simetría son los siguientes:

1. Se calcula la mediana de la variable (en nuestro caso 28,9).
2. Se ordenan los valores de la variable de mayor a menor (orden descendente).
3. Se calculan las diferencias d_i entre los valores de la variable ordenados y la mediana.
4. Se toman los valores positivos de d_i ordenados de menor a mayor y se les denomina p_i . Estos valores serán las *distancias sobre la mediana*.

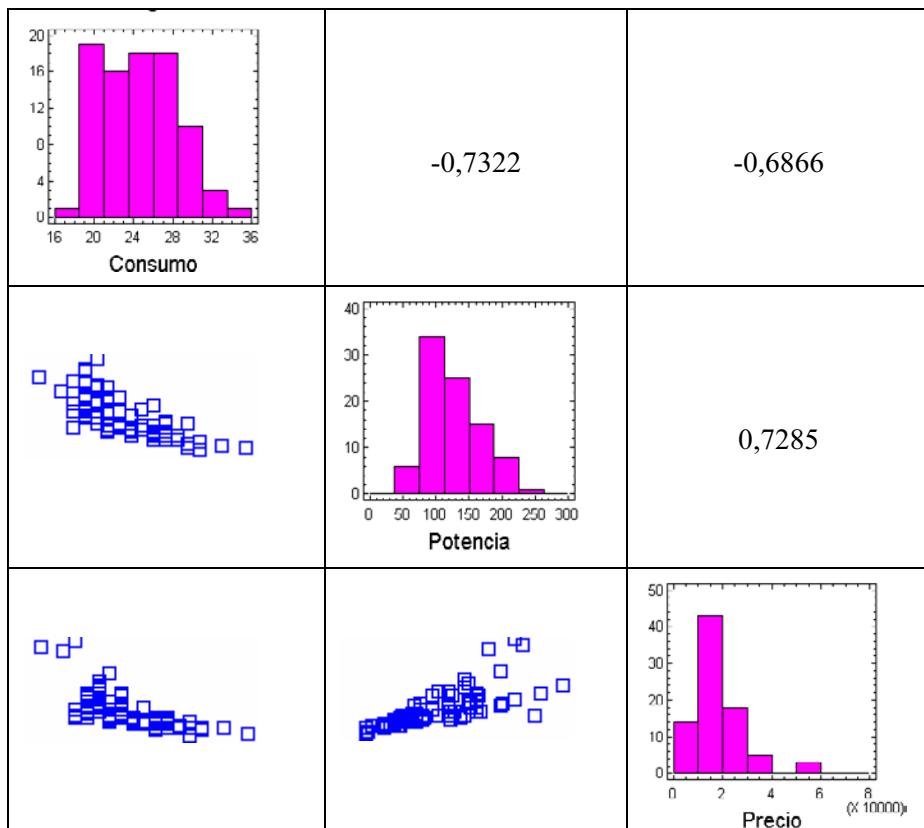
5. Se toman los valores negativos de d_i ordenados de menor a mayor y se les denomina n_i . Estos valores cambiados de signo serán las *distancias bajo la mediana*.
6. Se grafican los puntos de coordenadas $(-n_i, p_i)$.

En nuestro ejemplo comenzamos situando en una columna los valores de X ordenados de mayor a menor. Posteriormente formamos una segunda columna con los valores $X - 28,9$ resultantes de restar a X su mediana (28,9). A continuación partimos esta columna en dos columnas distintas. En la primera columna colocamos los valores negativos n_i de la columna $X - 28,9$ y los cambiamos de signo, y en la segunda columna colocamos los valores positivos p_i de la columna $X - 28,9$. El siguiente paso es ordenar ambas columnas (las dos ya positivas) de menor a mayor, resultando los valores para el grafo $(-n_i, p_i)$ de la siguiente Tabla:

$-n_i$	p_i	$-n_i$	p_i	$-n_i$	p_i
0,1	0,1	4,6	3,8	8,7	7,2
0,5	0,6	4,7	3,9	8,7	7,5
0,8	0,9	4,9	4	9	8,1
0,9	0,9	5	4,1	9,1	8,1
0,9	1	5	4,6	9,5	8,1
0,9	1,1	5,1	4,8	9,5	8,3
1	1,1	5,2	4,9	9,7	8,4
1,4	1,5	5,3	5,1	9,7	8,8
1,5	1,8	5,4	5,1	9,7	9,1
1,7	2	5,7	5,2	9,8	9,1
1,7	2,1	5,9	5,2	10,3	9,1
1,7	2,1	5,9	5,3	10,4	9,1
1,9	2,1	6,5	5,4	10,7	9,2
1,9	2,4	6,6	5,5	10,8	10,1
1,9	2,6	6,9	5,6	10,8	10,2
1,9	2,7	7,3	5,6	11,2	10,5
2,1	2,9	7,4	5,8	11,3	11,9
2,3	3	7,8	6,1	11,3	12
2,3	3,1	8,1	6,2	11,4	12,6
2,5	3,1	8,3	6,8	11,9	14,2
2,9	3,1	8,3	7,1	11,9	14,5
3,1	3,2	8,4	7,1	12	15,1
3,5	3,3	8,4	7,1	12,4	15,4
3,5	3,4	8,6	7,1	12,7	15,7
3,8	3,5	8,7	7,1	13,4	17,7
3,9	3,5	8,7	7,2		

Gráfico de dispersión

Se trata de un gráfico que permite ver la relación entre dos o más variables. Está formado por puntos cuyas coordenadas cartesianas son los pares de valores de dos variables cuya relación se quiere estudiar representada una en el eje vertical y otra en el eje horizontal. El posicionamiento de los puntos del gráfico de dispersión define la relación entre las variables. Si se sitúan alrededor de una recta, existe correlación lineal entre las variables. Si los puntos siguen una pauta no lineal, la relación entre las variables no puede definirse como lineal. Si la nube de puntos es aleatoria y dispersa, no existe relación alguna entre las variables. Un gráfico de dispersión para un análisis de varias variables que ofrece mucha información es el que representa en una estructura matricial los gráficos de dispersión de todos los pares posibles de variables (zona triangular inferior de la matriz), sus histogramas de frecuencias (diagonal de la matriz) y los coeficientes de correlación de todos los pares de variables en estudio (zona triangular superior de la matriz). En la Figura siguiente se muestra el gráfico matricial para las variables *Consumo*, *Potencia* y *Precio* de los automóviles.



El gráfico de dispersión del consumo con la potencia situado en el elemento (2,1) de la matriz es fácilmente ajustable a una recta de pendiente negativa, lo que indica una relación fuerte y negativa entre ambas variables. Esta información la corrobora el coeficiente de correlación entre consumo y potencia (-0,7322) situado en el elemento (1,2) de la matriz. El gráfico de dispersión del consumo con el precio situado en el elemento (3,1) de la matriz es fácilmente ajustable a una recta de pendiente negativa, lo que indica una relación fuerte y negativa entre ambas variables. Esta información la corrobora el coeficiente de correlación entre consumo y precio (-0,6866) situado en el elemento (1,3) de la matriz. El gráfico de dispersión de la potencia con el precio situado en el elemento (3,2) de la matriz es fácilmente ajustable a una recta de pendiente positiva, lo que indica una relación fuerte y positiva entre ambas variables. Esta información la corrobora el coeficiente de correlación entre potencia y precio (0,7285) situado en el elemento (2,3) de la matriz. Los histogramas de la diagonal principal de la matriz indican normalidad y simetría para el consumo y la potencia, y cierta asimetría para el precio.

Estadísticos robustos

Cuando se plantea el problema de analizar a qué estadísticos de centralización debe prestarse más atención a la hora de explorar los datos nos encontramos con que existen razones en pro y en contra para cada uno de ellos. La media aritmética es un estadístico que utiliza todos los datos de la variable en su elaboración, pero está muy afectada por los valores extremos. La mediana no resulta afectada por los valores extremos, pero su problema es que recoge información de muy pocos valores de la variable.

Se denominan *estadísticos robustos* aquellos que se ven poco afectados por la influencia de los valores extremos de la variable. La mediana es un estadístico de centralización robusto, pero la media no lo es. Nos encontramos así ante el problema de equilibrar entre la robustez de los estadísticos a utilizar como resumen de las variables y el número de observaciones afectadas por la definición de los estadísticos.

Una primera solución para el problema anterior es la consideración de la *media truncada*, que es la media de la variable eliminando el 5% de las colas inferior y superior de la distribución. De esta forma se elimina la influencia de los valores extremos a la vez que se incluyen en su cálculo el 90% de los valores centrales de la distribución.

Otra solución alternativa es la consideración de los *M-estimadores* caracterizados por su robustez al no verse afectados por los valores extremos. Estos estadísticos se definen ponderando cada valor en función de su distancia al centro de la distribución. Las observaciones centrales se ponderan por el máximo valor (la unidad) disminuyendo los coeficientes de ponderación a medida que las observaciones se alejan del centro de la distribución, llegando al extremo de ponderar con un cero valores muy lejanos al centro de la distribución (valores atípicos). La forma de ponderar clasifica los M-estimadores.

El *M-estimador de Hubert* pondera con el valor uno todos los valores situados a menos de 1,339 de la mediana. El *M-estimador de Tukey* pondera con un cero los valores situados a 4,385 de la mediana. El *M-estimador de Andrews* pondera con un cero los valores situados a 4,2066 de la mediana. El *M-estimador de Hampel* utiliza tres coeficientes de ponderación según que cada valor de la variable se encuentre a una distancia de la mediana de 1,7, 3,4 y 8,5 respectivamente.

Una propiedad importante de los estimadores robustos es que reducen notablemente la media de la distribución situándola muy cerca de la mediana. Se recomienda utilizar el M-estimador de Hubert cuando la distribución se acerca a la normalidad y no hay muchos valores extremos. Los M-estimadores de Tukey y Andrews son más útiles cuando existen casos atípicos.

Hasta aquí hemos analizado lo que podríamos denominar *estadísticos robustos centrales*. Pero también es necesario ocuparse de los *estadísticos robustos de dispersión*, que reflejan el grado en el que los datos tienden a extenderse alrededor del valor medio sin que haya demasiada influencia de los valores extremos.

Inicialmente se distingue entre medidas de dispersión absolutas y relativas, considerando relativas las que no dependen de las unidades de medida. Adicionalmente se clasifican las medidas absolutas y relativas según sean medidas referentes a promedios o no lo sean.

Entre las *medidas de dispersión absolutas no referentes a promedios* tenemos el *recorrido* o diferencia entre el mayor valor y el menor valor de una distribución, el *recorrido intercuartílico* o diferencia existente entre el tercer cuartil y el primero y el *recorrido semintercuartílico (desviación semintercuartil o amplitud intercuartil)* o recorrido intercuartílico dividido por dos. Es evidente que el recorrido no es un estadístico robusto, pues depende del mayor y menor valor de la variable, que precisamente son los valores extremos. Sin embargo, el recorrido intercuartílico y el recorrido semintercuartílico sí son estadísticos robustos porque en su definición no se incluyen el 25% superior e inferior de valores de la variable (no se incluyen los valores extremos).

Entre las *medidas de dispersión relativas no referentes a promedios* tenemos el *coeficiente de apertura* o cociente entre el mayor valor y el menor valor de una distribución, el *recorrido relativo* o cociente entre el recorrido y la media, así como el *coeficiente de variación intercuartílico* o cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil. Tanto el coeficiente de apertura como el recorrido relativo son estadísticos no robustos porque incluyen en su definición el máximo, el mínimo, el recorrido y la media, que dependen de valores extremos. Lo contrario ocurre con el coeficiente de variación intercuartílico, cuya definición excluye el 25% superior e inferior de valores de la variable (no depende de valores extremos).

Entre las *medidas de dispersión absolutas referentes a promedios* tenemos las *desviaciones medias*, la *varianza* y la *desviación típica*. Estas medidas de dispersión involucran a los promedios y permiten medir el error que cometemos utilizando el promedio en cuestión como resumen de los datos. En cuanto a las desviaciones medias tenemos en primer lugar la *desviación media respecto de la media aritmética*, que mide la eficacia de la media y que se define como la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la media aritmética, y cuya expresión es la siguiente:

$$D_m = \frac{1}{N} \sum_{i=1}^k |x_j - \bar{x}| n_i$$

Para medir la eficacia de la mediana Me suele considerarse la *desviación media respecto de la mediana*, que se define como la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la mediana, y cuya expresión es la siguiente:

$$D_{Me} = \frac{1}{N} \sum_{i=1}^k |x_j - Me| n_i$$

Tanto la varianza como la desviación típica y la desviación media respecto de la media involucran en su definición todos los valores de la variable, pero son estadísticos no robustos, porque su definición depende de la propia media que no es un estadístico robusto. La desviación media respecto de la mediana es un estadístico más robusto, por serlo la mediana, y además involucra en su definición todos los valores de la variable.

Entre las *medidas de dispersión relativas referentes a promedios* (valores adimensionales que no se ven afectados por las unidades de medida y que siempre se concretan en forma de cociente) tenemos el *índice de dispersión respecto a la mediana* y el *coeficiente de variación de Pearson*. El índice de dispersión respecto a la mediana se usa para resolver el problema de comparación de medianas de varias distribuciones que pueden venir, en general, en unidades diferentes y se define como la relación por cociente entre la desviación media respecto de la mediana y la mediana aritmética $V_{Me} = D_{Me} / Me$ (a menor índice de dispersión mejor es la mediana). El coeficiente de variación de Pearson se usa para resolver el problema de comparación de medias aritméticas de varias distribuciones que pueden venir, en general, en unidades diferentes y se define como la relación por cociente entre la desviación típica y la media aritmética $V = \sigma / \bar{x}$ (a menor coeficiente de variación V mejor es la media). Por otra parte, V representa el número de veces que σ contiene a \bar{x} , y es claro que cuanto mayor sea V más veces contendrá σ a \bar{x} , luego, relativamente, a mayor V menor representatividad de \bar{x} . Este coeficiente también se suele expresar en tantos por ciento como $V = 100 (\sigma / \bar{x})$.

Como tanto en el cálculo de σ como en el cálculo de \bar{x} han intervenido todos los valores de la distribución, V presenta la garantía, frente a otros coeficientes, de que utiliza toda la información de la distribución. La cota inferior de V es cero y el único caso problemático se presenta cuando $\bar{x} = 0$, lo que haría que V tendiera a infinito. Sin embargo, estamos ante un estadístico no robusto porque se ve muy afectado por los valores extremos de la distribución. La alternativa robusta al coeficiente de variación de Pearson es precisamente el coeficiente de variación intercuartílico definido anteriormente. Asimismo, también es un estadístico robusto el índice de dispersión respecto a la mediana, ya que su definición depende sólo de estadísticos robustos (la desviación media respecto a la mediana y la propia mediana).

Nos quedarían por analizar los *estadísticos robustos de asimetría y curtosis*. Las medidas de asimetría y curtosis suelen denominarse medidas de forma porque se basan en su representación gráfica sin llegar a realizar la misma. Las *medidas de asimetría* tienen como finalidad el elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presenta una distribución, sin necesidad de llevar a cabo su representación gráfica. Las *medidas de curtosis* estudian la distribución de frecuencias en la zona central de la misma. La mayor o menor concentración de frecuencias alrededor de la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada. Por esta razón a las medidas de curtosis se les llama también de apuntamiento o concentración central.

A continuación se definen las medidas de asimetría más comunes, entre las que destacan las siguientes:

Coefficiente de asimetría de Fisher: Su expresión es la siguiente:

$$g_1 = \frac{m_3}{\sigma^3} = \frac{\frac{1}{N} \sum_{i=1}^k (x_i - \bar{x})^3 n_i}{\left(\frac{1}{N} \sum_{i=1}^k (x_i - \bar{x})^2 n_i \right)^{3/2}}$$

Si $g_1 = 0$ la distribución es simétrica, si $g_1 > 0$ la distribución es asimétrica positiva (a derechas), y si $g_1 < 0$ la distribución es asimétrica negativa (a izquierdas). La distribución es asimétrica a derechas o positiva cuando la suma de las desviaciones positivas de sus valores respecto de la media es mayor que la suma de las desviaciones con signo negativo (la gráfica de la distribución tiene más densidad a la derecha de la media). En caso contrario, la distribución es asimétrica a la izquierda o negativa.

Coefficiente de asimetría de Fisher estandarizado: Para $N > 150$ el coeficiente de asimetría es asintóticamente normal de media cero y varianza $6/N$.

Este hecho nos lleva a considerar el coeficiente de asimetría estandarizado cuya expresión es:

$$g_s = \frac{g_1}{\sqrt{\frac{6}{N}}} \rightarrow N(0,1)$$

Este coeficiente es asintóticamente normal (0,1). Se trata por tanto de un estadístico que permite realizar el *contraste de asimetría formal* para una variable.

Coeficiente de asimetría de Pearson: Karl Pearson propuso para distribuciones campaniformes, unimodales y moderadamente asimétricas el coeficiente definido como $Ap = (\bar{x} - Mo) / \sigma$, donde Mo es la moda. Como en una distribución campaniforme simétrica $\bar{x} = Mo = Me$, si la distribución es asimétrica positiva o a derechas \bar{x} se desplaza a la derecha de la moda, y por tanto, $\bar{x} - Mo > 0$. En el caso de distribución asimétrica negativa la media se sitúa por debajo de Mo , por lo que el valor $\bar{x} - Mo < 0$. La desviación típica que aparece en el denominador no modifica el signo de la diferencia $\bar{x} - Mo$ y sirve para eliminar las unidades de medida de dicha diferencia. Así tendremos que si $Ap = 0$ la distribución es simétrica, si $Ap > 0$ la distribución es asimétrica positiva y si $Ap < 0$ la distribución es asimétrica negativa. También Pearson comprobó empíricamente para este tipo de distribuciones que se cumple $3(\bar{x} - Me) \approx \bar{x} - Mo$ (la mediana siempre se sitúa entre la media y la moda en las distribuciones moderadamente asimétricas). Por esta razón, algunos autores utilizan como coeficiente de asimetría de Pearson el valor $Ap \approx 3(\bar{x} - Me) / \sigma$.

El coeficiente absoluto de asimetría: Está basado también en la posición de los cuartiles y la mediana, y viene dado por la expresión:

$$A = [(C3 - C2) - (C2 - C1)] / S = (C3 + C1 - 2C2) / S = C3 + C1 - 2Me / S$$

Si $A = 0$ la distribución es simétrica, si $A > 0$ la distribución es asimétrica positiva y si $A < 0$ la distribución es asimétrica negativa. $C1$, $C2$ y $C3$ son los cuartiles de la distribución.

Coeficiente de asimetría de Bowley: Está basado en la posición de los cuartiles y la mediana, y viene dado por la expresión $Ab = (C3 + C1 - 2Me) / (C3 + C1)$. Se cumple que si $Ab = 0$ la distribución es simétrica, si $Ab > 0$ la distribución es asimétrica positiva y si $Ab < 0$ la distribución es asimétrica negativa. $C1$ y $C3$ son el primer y tercer cuartil respectivamente.

Índice de asimetría de Kelley: Se define como la diferencia entre la mediana y la semisuma de los deciles uno y nueve de la distribución ($Ik = Me - (D1+D9)/2$).

Tanto el coeficiente de asimetría de Pearson como el coeficiente absoluto de asimetría son estadísticos poco robustos porque su definición depende de magnitudes no robustas. Sin embargo el coeficiente de asimetría de Bowley y el índice de asimetría de Kelley son estadísticos robustos porque en su definición no intervienen las observaciones extremas.

Una vez presentadas las medidas de asimetría, a continuación se definen las medidas de curtosis más comunes, entre las que destacan las siguientes:

Coeficiente de curtosis: En la distribución normal se verifica que $m_4 = 3\sigma^4$, siendo m_4 el momento de orden 4 respecto a la media y σ la desviación típica. Si consideramos la expresión $g_2 = m_4/\sigma^4 - 3$, su valor será cero para la distribución normal. Por ello, como coeficiente de apuntamiento o curtosis se utiliza la expresión:

$$g_2 = \frac{m_4}{\sigma^4} - 3 = \frac{\frac{1}{N} \sum_{i=1}^k (x_j - \bar{x})^4 n_i}{\left(\frac{1}{N} \sum_{i=1}^k (x_j - \bar{x})^2 n_i \right)^2} - 3$$

Una distribución es *mesocúrtica* (apuntamiento igual al de la normal) cuando $g_2 = 0$, es *leptocúrtica* (apuntamiento mayor que el de la normal) si $g_2 > 0$, y es *platicúrtica* (apuntamiento menor que el de la normal) si $g_2 < 0$.

Coeficiente de curtosis estandarizado: Para $N > 150$ el coeficiente de curtosis es asintóticamente normal de media cero y varianza $24/N$. Este hecho nos lleva a considerar el coeficiente de curtosis estandarizado cuya expresión es:

$$g_{ks} = \frac{g_2}{\sqrt{\frac{6}{N}}} \rightarrow N(0,1)$$

Este coeficiente es asintóticamente normal (0,1). Se trata por tanto de un estadístico que permite realizar el contraste de curtosis formal para una variable.

Los coeficientes de curtosis definidos previamente no resultan ser estadísticos robustos porque su definición depende de estadísticos no robustos con fuerte influencia de los valores extremos.

Un índice de curtosis más robusto es el coeficiente de curtosis K2.

Transformaciones de las variables

Cuando el análisis exploratorio lo indique, los datos originales (no los estandarizados ni los previamente modificados) pueden necesitar ser transformados. Suelen considerarse cuatro tipos de transformaciones:

Transformaciones lógicas: Se unen categorías del campo de definición de las variables para reducir así su amplitud. De esta forma pueden eliminarse categorías sin respuestas. También pueden convertirse variables de intervalo en ordinales o nominales y crear variables ficticias (*dummy*).

Transformaciones lineales: Se obtienen al sumar, restar, multiplicar o dividir las observaciones originales por una constante para mejorar su interpretación. Estas transformaciones no cambian la forma de la distribución, ni las distancias entre los valores ni el orden, y por tanto no provocan cambios considerables en las variables.

Transformaciones algebraicas: Se obtienen al aplicar transformaciones no lineales monotónicas a las observaciones originales (raíz cuadrada, logaritmos, etc.) por una constante para mejorar su interpretación. Estas transformaciones cambian la forma de la distribución al cambiar las distancias entre los valores, pero mantienen el orden.

Transformaciones no lineales no monotónicas: Cambian las distancias y el orden entre los valores. Pueden cambiar demasiado la información original.

Con estas transformaciones se arreglan problemas en los datos. Por ejemplo: una asimetría negativa puede minorarse con una transformación parabólica o cúbica, una asimetría positiva fuerte puede suavizarse mediante una transformación hiperbólica o hiperbólica cuadrática (con signo negativo) y una asimetría positiva débil puede suavizarse mediante una transformación de raíz cuadrada, logarítmica o recíproca de la raíz cuadrada (con signo negativo). La transformación logarítmica puede conseguir estacionalidad en media y en varianza para los datos. Suele elegirse como transformación aquélla que arregla mejor el problema, una vez realizada. Si ninguna arregla el problema, realizamos el análisis sobre los datos originales sin transformar. Combinando transformaciones lineales y algebraicas pueden modificarse los valores extremos de la distribución.

ANÁLISIS DE LOS DATOS AUSENTES

Cuando se aplica un método de análisis multivariante sobre los datos disponibles puede ser que no exista información para determinadas observaciones y variables. Estamos entonces ante valores ausentes o valores missing. La presencia de esta información faltante puede deberse a un registro defectuoso de la información, a la ausencia natural de la información buscada o a una falta de respuesta (total o parcial).

Detección y diagnóstico de los datos ausentes

Tras observar la presencia de datos ausentes en una distribución será necesario detectar si estos se distribuyen aleatoriamente. Esta claro que el investigador debe averiguar si el proceso de ausencia de datos tiene lugar de forma aleatoria. La simple presencia de datos ausentes no implica que su falta sea crítica para el análisis estadístico. Será necesario detectar que el efecto de los datos ausentes es importante mediante pruebas formales de aleatoriedad.

Una primera prueba para valorar los datos ausentes para una única variable Y consiste en formar dos grupos de valores para Y, los que tienen datos ausentes y los que no los tienen. A continuación, para cada variable X distinta de Y, se realiza un test para determinar si existen diferencias significativas entre los dos grupos de valores determinados por la variable Y (ausentes y no ausentes) sobre X. Si vamos considerando como Y cada una de las variables del análisis y repitiendo el proceso anterior se encuentra que todas las diferencias son no significativas, se puede concluir que los datos ausentes obedecen a un *proceso completamente aleatorio* y por tanto pueden realizarse análisis estadísticos fiables con nuestras variables *imputando los datos ausentes* por los métodos que se verán más adelante. Si un porcentaje bastante alto de las diferencias son no significativas, puede considerarse que los datos ausentes obedecen a un *proceso aleatorio* (no completamente aleatorio) que también permitirá realizar análisis estadísticos fiables con nuestras variables previa **imputación de la información faltante**, aunque con menos fiabilidad que en el caso anterior.

Una segunda prueba para valorar los datos ausentes es la **prueba de las correlaciones dicotomizadas**. Para realizar esta prueba, para cada variable Y del análisis se construye una variable dicotomizada asignando el valor cero a los valores ausentes y el valor uno a los valores presentes. A continuación se dicotomizan todas las variables del análisis y se halla su matriz de correlaciones acompañada de los contrastes de significatividad de cada coeficiente de correlación de la matriz. Las correlaciones indican el grado de asociación entre los valores perdidos sobre cada par de variables (bajas correlaciones indican aleatoriedad en el par de variables), con lo que se puede concluir que si los elementos de la matriz de correlaciones son no significativos, los datos ausentes son completamente aleatorios. Si existe alguna correlación significativa y la mayor parte son no significativas, los datos ausentes pueden considerarse aleatorios. En ambos casos podrán realizarse análisis estadísticos previa imputación de la información faltante.

Una tercera prueba para valorar los datos ausentes es el **test conjunto de aleatoriedad de Little**, contraste formal basado en la chi-cuadrado, cuyo p-valor indica si los valores perdidos constituyen o no un conjunto de números aleatorios.

A continuación se ilustran los conceptos anteriores con un ejemplo basado en los datos recogidos en un cuestionario con 6 preguntas sobre comportamientos y actitudes de compra de 20 encuestados. Las respuestas a las 6 preguntas se recogen en 6 variables (V1, V2, V3, V4, V5 y V6) cuyo rango varía entre 1 y 10 reflejando la valoración que el encuestado da a la característica que refleja la pregunta. La primera pregunta valora la importancia que el encuestado da a la impresión que los demás tienen sobre él. La segunda pregunta refleja la valoración que el encuestado da a la garantía de las marcas. La tercera pregunta ofrece información sobre la frecuencia con que el encuestado compra sobre la marcha. La cuarta pregunta mide la preferencia que el encuestado da al comprar sobre ahorrar y vivir mejor. La quinta pregunta mide el gusto del encuestado por vestir a la moda y la quinta pregunta mide la tendencia del encuestado a conocer tiendas nuevas. Los datos de las 6 variables en los 20 cuestionarios se recogen en la Tabla siguiente:

Cuestionario	V1	V2	V3	V4	V5	V6
1	5	6	2	1	.	5
2	7	.	4	5	5	7
3	.	1	5	8	5	8
4	3	5	1	.	7	5
5	5	5	8	3	7	8
6	5	1	.	1	2	8
7	4	.	2	8	9	8
8	5	1	9	1	1	9
9	7	5	1	1	1	.
10	2	2	1	4	6	6
11	9	1	1	.	7	5
12	5	5	8	9	9	5
13	.	9	1	9	7	9
14	5	6	2	1	1	5
15	7	7	4	5	4	7
16	1	1	5	8	5	.
17	3	5	1	7	.	5
18	5	5	.	3	7	8
19	5	1	1	1	2	8
20	5	1	9	1	1	9

Una vez tabulada la información la primera tarea sería ver la tabla de frecuencias de los valores perdidos por variables para tener una idea de su magnitud. A continuación se presenta dicha información, observándose que para todas las variables el porcentaje de valores perdidos es del 10%, mientras que el de valores válidos es el 90%.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
V1	18	90,0%	2	10,0%	20	100,0%
V2	18	90,0%	2	10,0%	20	100,0%
V3	18	90,0%	2	10,0%	20	100,0%
V4	18	90,0%	2	10,0%	20	100,0%
V5	18	90,0%	2	10,0%	20	100,0%
V6	18	90,0%	2	10,0%	20	100,0%

El siguiente paso es determinar si los datos ausentes se distribuyen aleatoriamente. Para ello comparamos las observaciones con y sin datos ausentes para cada variable en función de las demás variables. La primera tarea será generar nuevas variables V11, V21, V31, V41, V51 y V61 (una para cada variable existente) asignándole el valor uno para datos válidos y el valor cero para datos ausentes. Tendremos la Tabla siguiente:

Cuest.	V1	V2	V3	V4	V5	V6	V11	V21	V31	V41	V51	V61
1	5	6	2	1	.	5	1	1	1	1	0	1
2	7	.	4	5	5	7	1	0	1	1	1	1
3	.	1	5	8	5	8	0	1	1	1	1	1
4	3	5	1	.	7	5	1	1	1	0	1	1
5	5	5	8	3	7	8	1	1	1	1	1	1
6	5	1	.	1	2	8	1	1	0	1	1	1
7	4	.	2	8	9	8	1	0	1	1	1	1
8	5	1	9	1	1	9	1	1	1	1	1	1
9	7	5	1	1	1	.	1	1	1	1	1	0
10	2	2	1	4	6	6	1	1	1	1	1	1
11	9	1	1	.	7	5	1	1	1	0	1	1
12	5	5	8	9	9	5	1	1	1	1	1	1
13	.	9	1	9	7	9	0	1	1	1	1	1
14	5	6	2	1	1	5	1	1	1	1	1	1
15	7	7	4	5	4	7	1	1	1	1	1	1
16	1	1	5	8	5	.	1	1	1	1	1	0
17	3	5	1	7	.	5	1	1	1	1	0	1
18	5	5	.	3	7	8	1	1	0	1	1	1
19	5	1	1	1	2	8	1	1	1	1	1	1
20	5	1	9	1	1	9	1	1	1	1	1	1

Ahora consideramos los dos grupos formados en la variable V1 (valores válidos y valores ausentes) que vienen definidos por la variable V11 y hacemos un contraste de igualdad de medias para los dos grupos de valores definidos en cada una de las restantes variables (V2 a V6) por los valores de V11. Tenemos el siguiente resultado:

V1	Prueba de Levene (para la igualdad de varianzas)		Prueba T para la igualdad de medias								
	F	Sig.	t	gl	Sig. (bilat.)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	Inferior	Superior	
V2	14,050	,002	,668	14	,515	1,36	2,033	-3,002	5,716		
			,335	1,048	,792	1,36	4,047	-44,817	47,532		
V3	,435	,520	-,321	14	,753	-,79	2,444	-6,028	4,456		
			-,360	1,412	,765	-,79	2,182	-15,118	13,546		
V4	3,168	,097	2,370	14	,033	4,93	2,079	,469	9,388		
			5,412	7,787	,001	4,93	,911	2,819	7,039		
V5	2,865	,113	,521	14	,610	1,14	2,192	-3,558	5,844		
			,894	2,595	,447	1,14	1,279	-3,311	5,597		
V6	4,359	,053	1,524	16	,147	1,75	1,148	-,684	4,184		
			2,753	2,549	,085	1,75	,636	-,492	3,992		

Se observa que salvo para la variable V4, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de V1 en las variables V2, V3, V5 y V6 (los intervalos de confianza contienen el valor cero). El contraste de igualdad de medias se realiza suponiendo varianzas iguales (primera línea de la tabla para cada variable) y desiguales (segunda línea para cada variable).

Ahora consideramos los dos grupos formados en la variable V2 (valores válidos y valores ausentes) que vienen definidos por la variable V21 y hacemos un contraste de igualdad de medias para los dos grupos de valores definidos en cada una de las restantes variables (V1 y V3 a V6) por los valores de V21. Tenemos el siguiente resultado:

V2	Prueba de Levene (varianzas)		Prueba T para la igualdad de medias								
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	Inferior	Superior	
V1	,290	,599	,439	14	,667	,57	1,300	-2,218	3,360		
			,365	1,188	,769	,57	1,566	-13,239	14,382		
V3	3,295	,091	-,321	14	,753	-,79	2,444	-6,028	4,456		
			-,587	3,067	,598	-,79	1,339	-4,995	3,424		
V4	1,160	,300	1,121	14	,281	2,64	2,358	-2,414	7,700		
			1,533	1,733	,283	2,64	1,724	-5,988	11,273		
V5	,309	,587	1,075	14	,301	2,29	2,127	-2,277	6,848		
			1,070	1,301	,444	2,29	2,137	-13,729	18,300		
V6	5,873	,028	,513	16	,615	,63	1,219	-1,959	3,209		
			,960	2,786	,413	,63	,651	-1,540	2,790		

Se observa que para todas las variables, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de V2 en cada una de ellas (los intervalos de confianza contienen el valor cero). Repitiendo los contrastes de igualdad de medias para los grupos que determinan los valores válidos y ausentes de las variables V3, V4, V5 y V6 en el resto de las variables, tenemos los siguientes resultados:

V3	Prueba de Levene (varianzas)		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
V1	1,956	,181	-,085	16	,933	-,13	1,473	-3,248	2,998
			-,246	15,000	,809	-,13	,507	-1,206	,956
V2	,187	,671	,409	16	,688	,81	1,988	-3,402	5,027
			,386	1,228	,756	,81	2,106	-16,631	18,256
V4	3,604	,076	1,048	16	,310	2,50	2,386	-2,559	7,559
			1,936	2,698	,158	2,50	1,291	-1,882	6,882
V5	,017	,898	,143	16	,888	,31	2,178	-4,305	4,930
			,120	1,169	,922	,31	2,600	-23,309	23,934
V6	9,655	,007	-,996	16	,334	-1,19	1,192	-3,715	1,340
			-2,893	15,000	,011	-1,19	,410	-2,062	-,313

V4	Prueba de Levene (varianzas)		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error tip. de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
V1	4,819	,043	-,868	16	,398	-1,25	1,440	-4,303	1,803
			-,413	1,038	,749	-1,25	3,028	-36,532	34,032
V2	,187	,671	,409	16	,688	,81	1,988	-3,402	5,027
			,386	1,228	,756	,81	2,106	-16,631	18,256
V3	5,206	,037	1,320	16	,206	2,94	2,226	-1,781	7,656
			3,833	15,000	,002	2,94	,766	1,304	4,571
V5	5,840	,028	-1,197	16	,249	-2,50	2,088	-6,926	1,926
			-3,478	15,000	,003	-2,50	,719	-4,032	-,968
V6	6,021	,026	1,988	16	,064	2,19	1,100	-,145	4,520
			5,775	15,000	,000	2,19	,379	1,380	2,995

V5	Prueba de Levene (varianzas)		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
V1	,054	,819	,689	16	,501	1,00	1,452	-2,079	4,079
			,897	1,536	,488	1,00	1,114	-5,490	7,490
V2	6,349	,023	-1,034	16	,317	-2,00	1,935	-6,102	2,102
			-2,405	6,336	,051	-2,00	,832	-4,009	,009
V3	3,618	,075	1,047	16	,310	2,38	2,268	-2,432	7,182
			2,565	8,439	,032	2,38	,926	,259	4,491
V4	,044	,837	,101	16	,921	,25	2,466	-4,978	5,478
			,080	1,148	,948	,25	3,106	-28,992	29,492
V6	6,021	,026	1,988	16	,064	2,19	1,100	-,145	4,520
			5,775	15,000	,000	2,19	,379	1,380	2,995

V6	Prueba de Levene (varianzas)		Prueba T para la igualdad de medias							
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	Inferior	Superior
V1	4,376	,053	,689	16	,501	1,00	1,452	-2,079	4,079	
			,330	1,039	,795	1,00	3,029	-34,218	36,218	
V2	,187	,671	,409	16	,688	,81	1,988	-3,402	5,027	
			,386	1,228	,756	,81	2,106	-16,631	18,256	
V3	,340	,568	,294	16	,772	,69	2,338	-4,268	5,643	
			,320	1,329	,792	,69	2,148	-14,850	16,225	
V4	,574	,460	-,127	16	,901	-,31	2,466	-5,540	4,915	
			-,087	1,103	,944	-,31	3,587	-36,945	36,320	
V5	,134	,719	,943	16	,360	2,00	2,121	-2,497	6,497	
			,943	1,264	,491	2,00	2,121	-14,691	18,691	

Se observa que para prácticamente todas las variables, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de cada una de ellas (los intervalos de confianza contienen el valor cero). Por lo tanto se puede concluir con bastante fiabilidad la distribución aleatoria de los dato perdidos, conclusión que permitirá realizar análisis estadísticos con los datos aplicando distintos métodos de imputación de la información faltante.

Para comprobar la aleatoriedad de los datos ausentes también se puede utilizar la matriz de correlaciones dicotomizadas. Se trata de calcular la matriz de correlaciones de las variables resultantes al sustituir los valores perdidos de las variables iniciales por ceros, y los valores válidos por unos. En nuestro caso se trataría de hallar la matriz de correlaciones de las variables V12 a V62. Tenemos los siguientes resultados:

		V11	V21	V31	V41	V51	V61
V11	Correlación de Pearson	1	-,111	-,111	-,111	-,111	-,111
	Sig. (bilateral)	.	,641	,641	,641	,641	,641
V21	Correlación de Pearson	-,111	1	-,111	-,111	-,111	-,111
	Sig. (bilateral)	,641	.	,641	,641	,641	,641
V31	Correlación de Pearson	-,111	-,111	1	-,111	-,111	-,111
	Sig. (bilateral)	,641	,641	.	,641	,641	,641
V41	Correlación de Pearson	-,111	-,111	-,111	1	-,111	-,111
	Sig. (bilateral)	,641	,641	,641	.	,641	,641
V51	Correlación de Pearson	-,111	-,111	-,111	-,111	1	-,111
	Sig. (bilateral)	,641	,641	,641	,641	.	,641
V61	Correlación de Pearson	-,111	-,111	-,111	-,111	-,111	1
	Sig. (bilateral)	,641	,641	,641	,641	,641	.

Las correlaciones resultantes entre las variables dicotómicas indican la medida en que los datos ausentes están relacionados entre pares de variables. Las correlaciones bajas indican una baja asociación entre los procesos de ausencia de datos para esas dos variables. En nuestro caso todas las correlaciones son bajas y significativas, lo que corrobora la presencia de aleatoriedad de los datos ausentes.

Soluciones para los datos ausentes: Supresión de datos o imputación de la información faltante

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico con ellos.

Podemos comenzar incluyendo sólo en el análisis las observaciones (casos) con datos completos (filas cuyos valores para todas las variables sean válidos), es decir, cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se denomina ***aproximación de casos completos o supresión de casos según lista*** y suele ser el método por defecto en la mayoría del software estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, porque su supresión provocaría una muestra representativa de la información total. En caso contrario se reduciría mucho el tamaño de la muestra a considerar para el análisis y no sería representativa de la información completa.

Otro método consiste en la ***supresión de datos según pareja***, es decir, se trabaja con todos los casos (filas) posibles que tengan valores válidos para cada par de variables que se consideren en el análisis independientemente de lo que ocurra en el resto de las variables. Este método elimina menos información y se utiliza siempre en cualquier análisis bivariante o transformable en bivariante.

Otro método adicional consiste en ***suprimir los casos (filas) o variables (columnas)*** que peor se comportan respecto a los datos ausentes. Nuevamente es necesario sopesar la cantidad de datos a eliminar. Debe siempre considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable o conjunto de casos en el análisis estadístico.

La alternativa a los métodos de supresión de datos es la ***imputación de la información faltante***. La imputación es el proceso de estimación de valores ausentes basado en valores válidos de otras variables o casos de la muestra. A continuación se estudian diferentes métodos de imputación.

Un primer método de imputación no reemplaza los datos ausentes sino que imputa las características de la distribución (por ejemplo, la desviación típica) o las relaciones de todos los valores válidos disponibles (por ejemplo, correlaciones).

El proceso de imputación no consiste en reemplazar los datos ausentes por el resto de los casos, sino en utilizar las características de la distribución o las relaciones de todos los valores válidos posibles, como representantes para toda la muestra entera. Este método se denomina ***enfoque de disponibilidad completa***.

Un segundo grupo de métodos de imputación ya son métodos de sustitución de datos ausentes por valores estimados sobre la base de otra información existente en la muestra. Consideraremos en este grupo el método de sustitución del caso, el método de sustitución por la media o la mediana, el método de sustitución por un valor constante, el método de imputación por interpolación lineal, el método de imputación por regresión y el método de imputación múltiple.

En el ***método de imputación por sustitución del caso*** las observaciones (casos) con datos ausentes se sustituyen con otras observaciones no maestrales. Por ejemplo, en una encuesta sobre hogares a veces se sustituye un hogar de la muestra que no contesta por otro hogar que no está en la muestra y que probablemente contestará. Este método de imputación suele utilizarse cuando existen casos con todas sus observaciones ausentes o con la mayoría de ellas.

En el ***método de imputación de sustitución por la media*** los datos ausentes se sustituyen por la media de todos los valores válidos de su variable correspondiente. Este método tiene la ventaja de que se implementa fácilmente y proporciona información completa para todos los casos, pero tiene la desventaja de que modifica las correlaciones e invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de los datos.

Cuando hay valores extremos en las variables, se sustituyen los valores ausentes por la mediana (en vez de por la media), ya que la mediana es un estadístico resumen de los datos más robusto. De esta forma se tiene el ***método de imputación de sustitución por la mediana***.

A veces, cuando hay demasiada variabilidad en los datos, suele sustituirse cada valor ausente por la media o mediana de un cierto número de observaciones adyacentes a él. En este tipo de imputación suele incluirse también el ***método de imputación por interpolación*** en el cual se sustituye cada valor ausente de una variable por el valor resultante de realizar una interpolación con los valores adyacentes.

En el ***método de imputación de sustitución por valor constante*** los datos ausentes se sustituyen por un valor constante apropiado derivado de fuentes externas o de una investigación previa. En este caso el investigador debe asegurarse de que la sustitución de los valores ausentes por el valor constante proveniente de una fuente externa es más válido que la sustitución por la media (valor generado internamente).

En el **método de imputación por regresión** se utiliza el análisis de la regresión para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos. Como desventaja de este método destacaríamos que refuerza las relaciones ya existentes en los datos de modo que conforme aumente su uso los datos resultantes son más característicos de la muestra y menos generalizables. Además, con este método se subestima la varianza de la distribución. Y no olvidemos como desventaja que este método supone que la variable con datos ausentes tiene correlaciones sustanciales con otras variables.

El **método de imputación múltiple** es una combinación de varios métodos de entre los ya citados.

ANÁLISIS Y DETECCIÓN DE VALORES ATÍPICOS

Los casos atípicos son observaciones aisladas cuyo comportamiento se diferencia claramente del comportamiento medio del resto de las observaciones.

Existe una primera categoría de casos atípicos formada por aquellas observaciones que provienen de un error de procedimiento, como por ejemplo un error de codificación, error de entrada de datos, etc. Estos datos atípicos, si no se detectan mediante filtrado, deben eliminarse o recodificarse como datos ausentes.

Una segunda categoría de casos atípicos contempla aquellas observaciones que ocurren como consecuencia de un acontecimiento extraordinario existiendo una explicación para su presencia en la muestra. Este tipo de casos atípicos normalmente se retienen en la muestra, salvo que su significancia sea sólo anecdótica.

Una tercera categoría de datos atípicos comprende las observaciones extraordinarias para las que el investigador no tiene explicación. Normalmente estos datos atípicos se eliminan del análisis.

Una cuarta categoría de casos atípicos la forman las observaciones que se sitúan fuera del rango ordinario de valores de la variable. Suelen denominarse valores extremos y se eliminan del análisis si se observa que no son elementos significativos para la población.

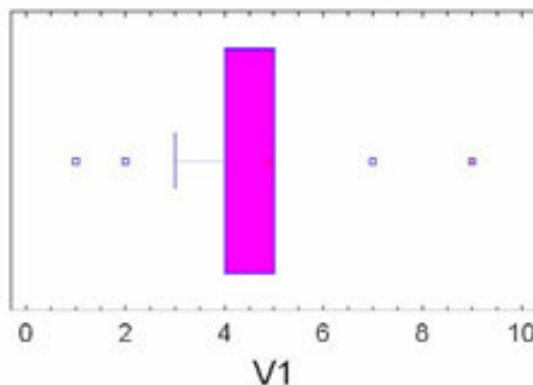
Las propias características del caso atípico, así como los objetivos del análisis que se realiza, determinan los casos atípicos a eliminar.

No obstante, los casos atípicos deben considerarse en el conjunto de todas las variables consideradas. Por lo tanto, hay que analizarlos desde una perspectiva multivariante. Puede ocurrir que una variable tenga valores extremos eliminables, pero al considerar un número suficiente de otras variables en el análisis, el investigador puede decidir no eliminarlos.

Detección univariante de casos atípicos

Cuando se trata de detectar casos atípicos en un contexto univariante, pueden utilizarse *herramientas de análisis exploratorio de datos*, por ejemplo el Gráfico de Caja y Bigotes. En este gráfico los valores atípicos se presentan como puntos aislados en los extremos de los bigotes. Los valores extremos suelen aparecer tachados con una *x*. El software habitual indica el número de observación correspondiente a los valores atípicos. En el ejemplo siguiente se muestra el Gráfico de Caja y Bigotes para la variable V1 ya definida anteriormente.

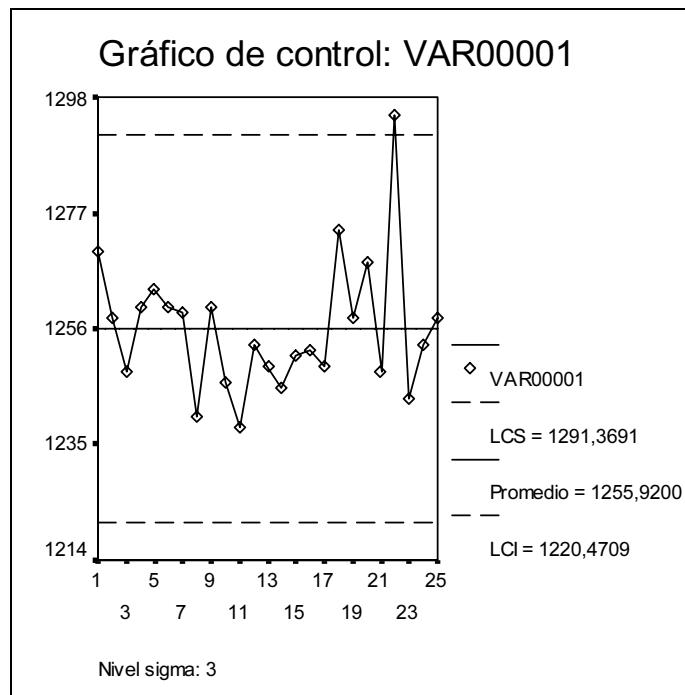
Gráfico de Caja y Bigotes



Se observan dos valores atípicos anteriores al bigote izquierdo y otros dos posteriores al bigote derecho. El último de ellos es un valor extremo (aparece tachado).

También pueden detectarse los valores atípicos de una variable mediante un *diagrama de control*, que es una representación gráfica con una línea central que representa el valor medio de la variable y con otras dos líneas horizontales, llamadas Límite Superior de Control (LSC) y Límite Inferior de Control (LIC). Se escogen estos límites de manera que casi la totalidad de los puntos de la variable se halle entre ellos. Mientras los valores de la variable se encuentran entre los límites de control, se considera que no hay valores atípicos. Sin embargo, si un punto que se encuentra fuera de los límites de control se interpreta como un valor atípico, y son necesarias acciones de investigación y corrección a fin de encontrar y eliminar la o las causas asignables a este comportamiento. Se acostumbra a unir los diferentes puntos en el diagrama de control mediante segmentos rectilíneos con objeto de visualizar mejor la evolución de la secuencia de los valores de la variable. Sin importar la distribución de la variable, es práctica estándar situar los límites de control como un múltiplo de la desviación típica. Se escoge en general el múltiplo 3, es decir, se acostumbra utilizar los límites de control de tres sigmas en los diagramas de control.

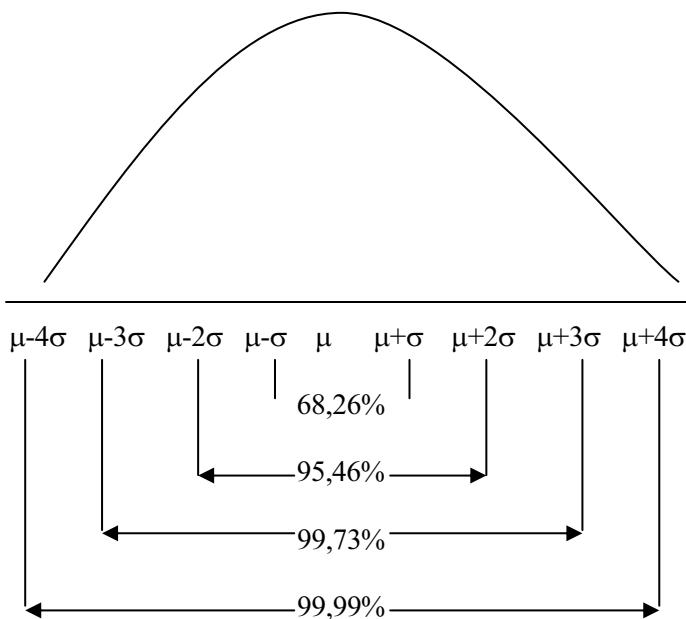
A continuación se presenta el gráfico de control tres sigmas para una variable con los 25 valores entre 1 238 y 1 295 siguientes: 1 270, 1 258, 1 248, 1 260, 1 263, 1 260, 1 259, 1 240, 1 260, 1 246, 1 238, 1 253, 1 249, 1 245, 1 251, 1 252, 1 249, 1 274, 1 258, 1 268, 1 248, 1 295, 1 243, 1 253, 1 258,00



Se ve que la observación número 22 es un valor atípico por caer fuera de los límites de control.

La razón fundamental del uso de límites tres sigmas radica en que la mayoría de las distribuciones con que nos encontramos normalmente se aproximan a la forma de campana de Gauss correspondiente a la función de densidad de la distribución normal. Si usamos la desviación estándar (sigma) para dividir el área que se encuentra debajo la curva, tal como se indica en la Figura de la página siguiente, podemos calcular las áreas de cada zona limitada por los valores $\mu \pm k\sigma$ de la abscisa como un porcentaje del área total que hay debajo de las curvas. Como indica la Figura, la probabilidad de encontrar un valor dentro de $\mu \pm \sigma$ es aproximadamente del 68%, o lo que es lo mismo; la probabilidad de obtener un valor fuera de estos límites es aproximadamente del 32%. Similarmente, la probabilidad de que los valores caigan fuera de los límites $\mu \pm 2\sigma$ es aproximadamente del 4,5%, mientras que la probabilidad de que los valores caigan fuera de los límites $\mu \pm 3\sigma$ es ya pequeñísima (sólo del 0,3% o del tres por mil).

Puesto que un acontecimiento que tenga esta probabilidad tan baja sucede muy raramente, usualmente cuando los datos caen fuera de los límites 3-sigma, sacamos la conclusión de que la distribución ha cambiado, de que el proceso definido por la variable ha cambiado, y de que presenta alguna anomalía. Ésta es la justificación del uso generalizado de los límites tres sigmas para detectar los valores atípicos.



Otra forma de detectar la existencia de posibles valores atípicos es utilizar los *estadísticos robustos de la variable* y ver su diferencia respecto de los estadísticos no robustos. Suelen considerarse como estadísticos robustos de centralización (localización) la mediana, la media truncada y la media winsorizada. La media truncada prescinde del 15% de los valores de la variable por cada extremo y la media winsorizada sustituye ese 15% de valores por valores del centro de la distribución. Como estadísticos robustos de dispersión (escala) se usan respectivamente la variación media respecto de la mediana, la desviación típica truncada y la desviación típica winsorizada. Cuando no hay valores atípicos, los estadísticos robustos y los estadísticos normales no difieren mucho. También pueden calcularse intervalos de confianza para la media normal y para la media winsorizada. Si su anchura es similar no hay valores atípicos.

No obstante, es más efectivo utilizar un *contraste formal estadístico para detectar valores atípicos*, por ejemplo el *test de Dixon* o el *test de Grubbs*, cuyos p-valores detectan los valores atípicos. Para p-valores menores que 0,05 hay valores atípicos al 95% de confianza.

A continuación se muestra la detección de los valores atípicos para la variable *var0001* con 25 valores entre 1 238 y 1 295 analizada anteriormente.

Identificación de valores atípicos

Datos: var00001

25 valores comprendidos desde 1238,0 hasta 1295,0

Número de valores actualmente excluidos: 0

Estimación de la localización:

Media de la Muestra = 1255,92

Mediana de la Muestra = 1253,0

Media Truncada = 1254,47

Media Winsorizada = 1255,08

Estimación de la escala:

Desv. Típica de la muestra = 12,1755

MAD/0.6745 = 10,3781

Sbi = 10,3128

Sigma Winsorizada = 9,25515

95,0% intervalos de confianza para la media:

Estándar: (1250,89,1260,95)

Winsorizada: (1250,62,1259,54)

Truncada: 15,0%

Valores ordenados

Fila	Valor	Valores Estudentizados Sin supresión	Valores Estudentizados Con supresión	Modificados MAD puntuación Z
11	1238,0	-1,47181	-1,57681	-1,44536
8	1240,0	-1,30754	-1,38575	-1,25264
23	1243,0	-1,06115	-1,10954	-0,963571
14	1245,0	-0,896881	-0,930977	-0,770857
10	1246,0	-0,814749	-0,843062	-0,6745
...				
5	1263,0	0,581495	0,59737	0,963571
20	1268,0	0,992155	1,03407	1,44536
1	1270,0	1,15642	1,21503	1,63807
18	1274,0	1,48495	1,59237	2,0235
22	1295,0	3,20972	4,402	4,047

Test de Grubbs (asume normalidad)

Test estadístico = 3,20972

p-valor = 0,00646216

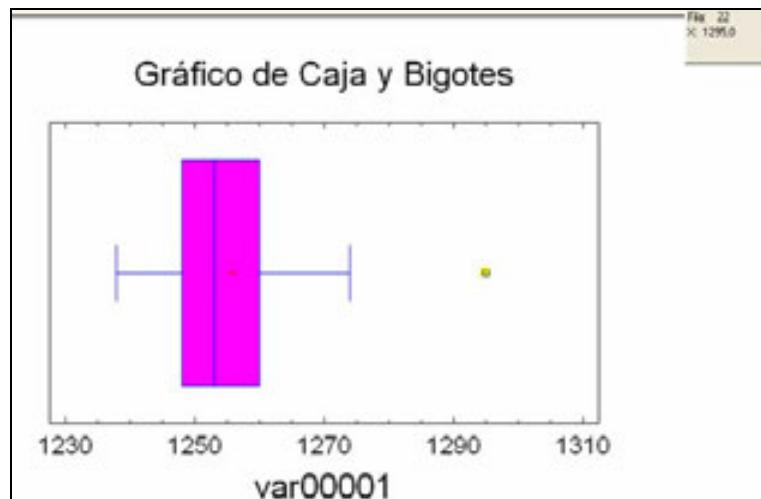
Test de Dixon (asume normalidad)

	Estadístico	5% Test	1% Test
1 valor atípico a la derecha	0,381818	Significativo	No sig.
1 valor atípico a la izquierda	0,0555556	No sig.	No sig.
2 valores atípicos a la derecha	0,454545	Significativo	No sig.
2 valores atípicos a la izquierda	0,138889	No sig.	No sig.
1 valor atípico en cualquier lado	0,368421	Significativo	No sig.

Este análisis identifica y trata los potenciales valores atípicos en muestras procedentes de poblaciones normales. En la parte superior de la salida se muestran las estimaciones usuales de la media y la desviación típica, junto con las estimaciones robustas resistentes a los valores atípicos. Para los 25 valores de *var00001*, la media y sigma son 1 255,92 y 12,1755 respectivamente (9,7 % de coeficiente de variación). La mediana y la desviación media respecto de la mediana son 1 253 y 10,3781 respectivamente (8,2% índice de variación respecto de la mediana). La media truncada y la desviación típica truncada son 1 254,47 y 10,3128 (8,2 de coeficiente de variación truncado). La media winsorizada y la desviación típica winsorizada son 1 255,08 y 9,25515 (7,4 de coeficiente de variación winsorizado). Se advierte el impacto de las estimaciones winsorizadas en el coeficiente de variación y en intervalo de confianza para la media, por lo que probablemente existan valores atípicos.

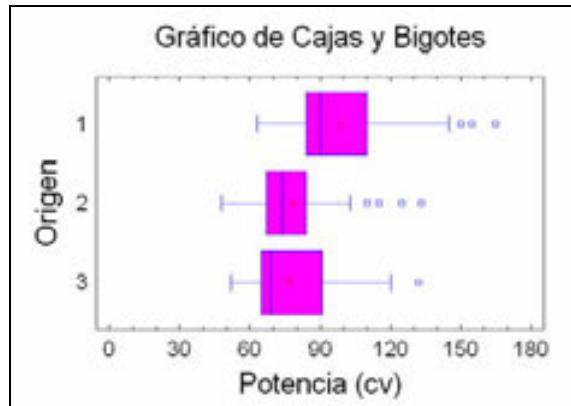
La parte central de la salida muestra los valores de *var00001* ordenados de menor a mayor. Los valores estudiantizados miden cuántas desviaciones típicas de cada valor se desvían de la media sin supresión 1 255,92 y con supresión de cada punto uno a uno, así como cuando la media y la desviación típica están basadas en la desviación respecto de la mediana (MAD). El valor más extremo es el de la fila 22, que es 3,20972 desviaciones típicas de la media. Dado que el p-valor para el *test de Grubbs* relativo a este valor más extremo es inferior a 0,05, ese valor es un atípico significativo al 95% de confianza, asumiendo que el resto de los valores siguen una distribución normal. Valores de los resultados basados en MAD superiores a 3,5 en valor absoluto, de los cuales hay 1, pueden ser fácilmente valores atípicos. También se ha realizado el *test de Dixon*, que en este caso, indica 1 valor atípico significativo en cualquier lado.

Si observamos el gráfico de Caja y Bigotes para *var00001* se ve que efectivamente el valor 1 295 que ocupa la fila 22 es efectivamente atípico.

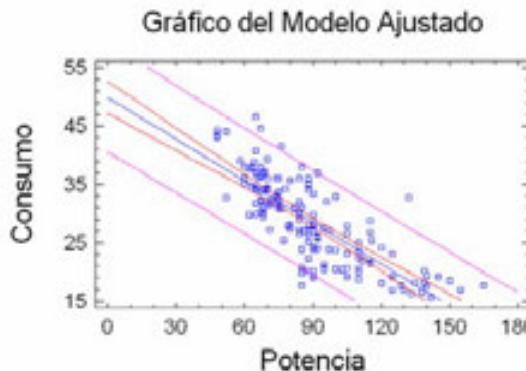


Detección bivariante de casos atípicos

Cuando se trata de detectar casos atípicos en un contexto bivariante, pueden utilizarse *herramientas de análisis exploratorio de datos*, por ejemplo el Gráfico de Caja y Bigotes múltiple que representa distintos gráficos de una variable (potencia de los automóviles) para diferentes niveles de la otra (país de origen). Se observan valores atípicos para los tres orígenes (3 para el origen uno, 4 para el dos y 1 para el tres).



También pueden evaluarse conjuntamente pares de variables mediante un gráfico de dispersión. Casos que caigan manifiestamente fuera del rango del resto de las observaciones pueden identificarse como puntos aislados en el gráfico de dispersión. La delimitación de los puntos aislados puede reflejarse mediante una elipse que represente el intervalo de confianza al 95% para una distribución normal bivariante, o mediante bandas de confianza del 95% tal y como se observa en la Figura siguiente que representa el consumo de los coches en función de su potencia. Se observan 5 valores atípicos por encima de la banda de confianza y 3 por debajo.



Detección multivariante de casos atípicos

Cuando se trata de detectar casos atípicos en un contexto multivariante, pueden utilizarse *estadísticos basados en distancias*, para detectar los puntos influyentes. La *distancia D² de Mahalanobis* es una medida de la distancia de cada observación en un espacio multidimensional respecto del centro medio de las observaciones. El *estadístico DFITS* mide la influencia de cada observación en caso de ser eliminada del análisis. La *Influencia (Leverage)* mide la influencia de cada observación. Por ejemplo, si se consideran las variables potencia, cilindrada y consumo de los automóviles, los estadísticos *D² de Mahalanobis*, *DFITS* e *Influencia* se presentan a continuación:

Puntos Influyentes			
Fila	Influencia	Distancia de Mahalanobis	DFITS
20	0,1153550	18,3055	0,424114
38	0,0220039	2,33655	-0,294588
57	0,0916601	13,9413	-0,121362
79	0,0133516	1,00948	0,330362
82	0,0330894	4,07154	0,283416
90	0,0327292	4,01454	0,579316
122	0,0621604	8,81621	0,237141
124	0,0220039	2,33655	-0,372586
145	0,0220039	2,33655	0,328165

Influencia media de un punto = 0,02

En este caso, un punto medio tendría un valor de influencia igual a 0,02. Hay 3 puntos superiores a 3 veces la influencia media (filas 20, 57 y 122), uno superior a 5 veces (fila 20). Deberían examinarse cuidadosamente los puntos superiores a 5 veces la influencia media (observación número 20) y determinar cuánto cambiaría el modelo si éstos se eliminaran. Además, hay 7 puntos con un valor DFITS extraordinariamente alto.

COMPROBACIÓN DE LOS SUPUESTOS DEL ANÁLISIS MULTIVARIANTE

Una etapa importante en el análisis multivariante es la comprobación de los supuestos estadísticos subyacentes a las variables que intervienen en los modelos. La presencia de múltiples variables provoca complejidad de relaciones que llevan a distorsiones y sesgos cuando no se cumplen determinados supuestos que se estudiarán a continuación (normalidad, homoscedasticidad, linealidad, ausencia de autocorrelación o correlación serial y ausencia de multicolinealidad).

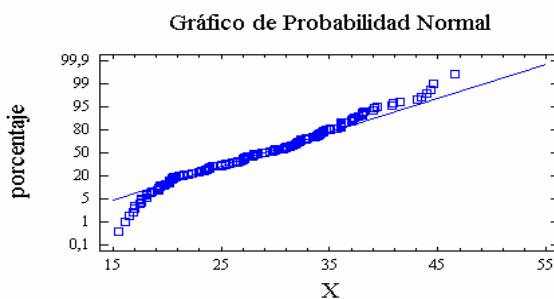
Normalidad

Tanto los métodos estadísticos univariantes como los multivariantes se basan en los supuestos de normalidad univariante y multivariante respectivamente. Todas las variables que intervienen en un método de análisis multivariante deben ser normales, aunque ello no garantiza la normalidad multivariante. El recíproco siempre es cierto, es decir, la normalidad multivariante implica la normalidad de cada variable. No obstante suele bastar con la normalidad de cada variable, aunque en procesos críticos puede exigirse la normalidad multivariante.

Existen, tanto métodos gráficos, como contrastes estadísticos formales, para comprobar la normalidad de las variables que intervienen en un método multivariante.

Gráfico normal de probabilidad

Los gráficos normales de probabilidad sirven para determinar si un conjunto de datos dado se ajusta razonablemente a una distribución normal. El gráfico normal de probabilidad presenta en el eje de abscisas los valores de la variable (X_i), y en el eje de ordenadas las frecuencias relativas acumuladas de dichos valores (F_i). La normalidad de los datos será perfecta cuando el gráfico de los puntos (X_i, F_i) resulte ser una línea recta situada sobre la diagonal del primer cuadrante. Las diferencias que existan entre el gráfico de probabilidad y la línea recta marcarán la regla de decisión para aceptar o no la normalidad del conjunto de datos dado.



Se observa que la variable X se ajusta bastante bien a una normal, ya que los puntos de la gráfica se aproximan bastante a la diagonal.

La Tabla de frecuencias resumida para la variable X con valores sin agrupar que permite realizar el gráfico (X_i, F_i) de normalidad es la siguiente:

<i>Nº Observ.</i>	<i>X_i</i>	<i>n_i</i>	<i>f_i</i>	<i>N_i</i>	<i>F_i</i>
1	15,50	1	0,0065	1	0,0065
2	16,20	1	0,0065	2	0,0130
3	16,50	1	0,0065	3	0,0195
4	16,90	1	0,0065	4	0,0260
5	17,00	2	0,0130	6	0,0390
6	17,50	1	0,0065	7	0,0455
7	17,60	2	0,0130	9	0,0584
8	17,70	1	0,0065	10	0,0649
9	18,10	2	0,0130	12	0,0779
10	18,20	1	0,0065	13	0,0844
11	18,50	1	0,0065	14	0,0909
12	18,60	1	0,0065	15	0,0974
13	19,10	1	0,0065	16	0,1039
14	19,20	3	0,0195	19	0,1234
15	19,40	2	0,0130	21	0,1364
16	19,80	1	0,0065	22	0,1429
17	19,90	1	0,0065	23	0,1494
18	20,20	4	0,0260	27	0,1753
19	20,30	1	0,0065	28	0,1818
20	20,50	2	0,0130	30	0,1948
21	20,60	2	0,0130	32	0,2078
22	20,80	1	0,0065	33	0,2143
23	21,10	1	0,0065	34	0,2208
24	21,50	1	0,0065	35	0,2273
25	21,60	1	0,0065	36	0,2338
26	22,00	1	0,0065	37	0,2403
.
.
100	38,10	1	0,0065	142	0,9221
101	39,00	1	0,0065	143	0,9286
102	39,10	1	0,0065	144	0,9351
103	39,40	1	0,0065	145	0,9416
104	40,80	1	0,0065	146	0,9481
105	40,90	1	0,0065	147	0,9545
106	41,50	1	0,0065	148	0,9610
107	43,10	1	0,0065	149	0,9675
108	43,40	1	0,0065	150	0,9740
109	44,00	1	0,0065	151	0,9805
110	44,30	1	0,0065	152	0,9870
111	44,60	1	0,0065	153	0,9935
112	46,60	1	0,0065	154	1,0000

Contrastes de la bondad de ajuste: test de la chi-cuadrado

Este tipo de tests trata de contrastar si de los datos obtenidos en una muestra se puede deducir o no que proceden de una población con una distribución determinada (por ejemplo, de la distribución normal). Entre ellos destaca el contraste chi-cuadrado de la bondad de ajuste que se describe a continuación.

Sea una muestra aleatoria simple de tamaño N de una cierta variable. El contraste de la chi-cuadrado de la bondad de ajuste se lleva a cabo para saber si hay razones suficientes para aceptar que esta variable tiene una distribución de probabilidad dada. Por tanto, la hipótesis nula H_0 hará referencia a que las observaciones muestrales constituyen un conjunto de N valores procedentes de una variable aleatoria con una distribución de probabilidad dada por $P(S)$, y la alternativa de que no sucede esto. Pasemos a estudiar este contraste distinguiendo dos casos:

a) Parámetros poblacionales conocidos

La distribución de probabilidad $P(S)$ está completamente especificada, de modo que en su expresión no aparecen parámetros desconocidos. Si H_0 es cierta, la distribución de la muestra, que es la distribución uniforme discreta que se obtiene al asignar una probabilidad $P_i = 1/N$ a cada uno de los N valores observados, puede considerarse como una imagen estadística de la distribución poblacional $P(S)$.

Sin embargo, debido a fluctuaciones aleatorias del muestreo, lo razonable será que las distribuciones poblacional y muestral no coincidan, pero para muestras grandes cabe esperar que la distribución de la muestra constituya una aproximación a la distribución poblacional. En este sentido, parece natural introducir alguna medida de la desviación entre ambas distribuciones y basar nuestro contraste en las propiedades de la distribución en el muestreo de esa medida.

Supongamos que el espacio de la variable aleatoria se divide en r intervalos S_1, S_2, \dots, S_r disjuntos, y sea la probabilidad de que un valor pertenezca al intervalo i -ésimo $p_i = P(S_i)$, donde tales intervalos pueden ser los r grupos o clases en que se han dispuesto los valores muestrales a efectos de tabulación. Sean n_1, n_2, \dots, n_r las correspondientes frecuencias absolutas en los r grupos muestrales, de tal manera que n_i valores muestrales pertenecen a la clase S_i , siendo $\sum n_i = N$ con $i = 1, \dots, r$. Ahora se elabora una medida de la desviación entre la distribución de la muestra y la distribución supuesta para la población bajo la hipótesis nula, cuya expresión es:

$$\chi^2 = \sum \frac{(n_i - Np_i)^2}{Np_i}$$

donde n_i son las frecuencias observadas en la muestra, y Np_i la estimación de las frecuencias esperadas si la distribución poblacional fuese la que indica H_0 . El estadístico χ^2 tiene, en el límite ($N \rightarrow \infty$), una distribución chi-cuadrado con $r - 1$ grados de libertad.

Determinada la distribución en el muestreo para la medida de desviación χ^2 , construiremos el contraste para H_0 de la siguiente forma: fijado el nivel de significación α , buscamos en las tablas de la chi-cuadrado un valor C tal que $P(\chi^2_{(r-l, \alpha)} > C) = \alpha$, y si el valor del estadístico χ^2 es mayor que C , entonces rechazamos la hipótesis H_0 de que la muestra proviene de la población indicada. Si el valor del estadístico χ^2 es menor que C , entonces aceptamos H_0 y aseguramos que la muestra proviene de la población indicada.

b) Parámetros poblacionales desconocidos

Supongamos que según la hipótesis nula H_0 la muestra procede de una población en donde hay que estimar previamente una serie de k parámetros por ser desconocidos. Fisher demostró que si previamente se estiman los k parámetros mediante la información muestral, el límite de la distribución del estadístico χ^2 es una chi-cuadrado con $r - k - 1$ grados de libertad, y el contraste se realiza como en el caso anterior. La única diferencia son los grados de libertad del estadístico χ^2 .

Una vez expuesto el contraste χ^2 de bondad del ajuste, tanto para el caso en que no se estiman parámetros como para cuando sí se hace, hemos de señalar que, para que no se produzcan perturbaciones, los intervalos S_i deberían ser tales que los valores de $P(S_i)$, para todo i , fueran aproximadamente iguales. Además, cuando el número de intervalos en que se divide el espacio de la variable aleatoria sea 2 el test χ^2 no debe utilizarse si existe alguna frecuencia esperada inferior a 5.

Si el número de intervalos es mayor que 2 tampoco deberá usarse el test χ^2 si más del 20% de las frecuencias esperadas son menores que 5 o alguna es inferior a 1. No obstante, las frecuencias esperadas a veces pueden incrementarse realizando uniones de intervalos adyacentes, uniones que, por supuesto, tengan sentido en la realidad.

Contraste de Kolmogorov-Smirnov Lilliefors de la bondad de ajuste

El contraste de Kolmogorov-Smirnov es un test para bondad de ajuste alternativo al de la chi-cuadrado. Al igual que en el contraste de la chi-cuadrado de bondad del ajuste, consideramos que la masa total de probabilidad discreta está repartida uniformemente entre los N valores muestrales de forma que, ordenados los valores muestrales de menor a mayor, la función de distribución empírica de la muestra es $F_n(x) = N_i/N$.

El contraste de Kolmogorov-Smirnov se aplica solo a variables continuas y trata de medir el ajuste entre la función de distribución empírica de una muestra y la función de distribución teórica. Se trata por tanto de un contraste de ajuste de la distribución de una muestra dada a una distribución continua determinada.

La función de distribución empírica de una muestra x_1, x_2, \dots, x_n , se define como:

$$F_n(x) = \frac{\text{nº de valores del conjunto } \{x_1, x_2, \dots, x_n\} \text{ que son } \leq x}{n}$$

Para contrastar la hipótesis de que la muestra se ajusta a una distribución teórica $F(x)$, se calcula el estadístico:

$$D_n = \text{Máx } |F_n(x) - F(x)|$$

cuya distribución es conocida y está tabulada. Si la distancia calculada D_n es mayor que la encontrada en las tablas, para un nivel α , rechazamos la distribución $F(x)$ para la muestra.

Para n y α dados, hallamos $D(\alpha, n)$ tal que $P(D_n > D(\alpha, n)) = \alpha$. La región crítica del test será $D_n > D(\alpha, n)$.

Este contraste tiene la ventaja de que no requiere agrupar los datos y el inconveniente de que si calculamos $F(x)$ estimando parámetros de la población, mediante la muestra, la distribución de D_n es sólo aproximada.

Normalmente se usa el estadístico $\sqrt{n} D_n$ en vez del D_n , lo que nos permite contrastar tablas para tamaños muestrales muy grandes.

Tendremos en cuenta las siguientes consideraciones:

- El test de Kolmogorov-Smirnov es de más fácil aplicación que el test de la χ^2 .
- El test de Kolmogorov-Smirnov no se ve afectado por reagrupaciones de las observaciones, mientras que en el test de la χ^2 , al disminuir los grupos, se pierde información, así como grados de libertad.
- El test de Kolmogorov-Smirnov es aplicable a pequeñas muestras, mientras que el test de la χ^2 está diseñado para grandes muestras.
- La potencia del test de Kolmogorov-Smirnov es mayor que la del de la χ^2 , si bien tienden a igualarse cuando el tamaño de la muestra crece.
- El test de la χ^2 puede fácilmente ser modificado cuando hay parámetros desconocidos mientras que el test de Kolmogorov-Smirnov no tiene tal flexibilidad.
- El test de la χ^2 es aplicable cuando la población es discreta o continua y el test de Kolmogorov-Smirnov requiere la continuidad de $F(x)$.

Se puede utilizar para estos contrastes (como para todos) el criterio del p-valor, rechazando la hipótesis nula al nivel α cuando el p-valor es menor que α , y aceptándola en caso contrario.

Cuando la distribución a ajustar es una normal, el estadístico de Kolmogorov-Smirnov fue estudiado y corregido por Lilliefors.

Contraste de normalidad de Shapiro y Wilks

Los contrastes de normalidad son un caso particular de contraste de ajuste, donde se trata de comprobar si los datos provienen de una distribución normal. Además del contraste de la chi-cuadrado y el contraste de Kolmogorov-Smirnov ya estudiados, que sirven para hacer ajustes a cualquier distribución (incluida la normal) existen otros contrastes específicos para normalidad como el contraste de normalidad W de Shapiro y Wilks, y los contrastes Z de asimetría y curtosis.

El contraste de Shapiro y Wilks mide el ajuste de la muestra a una recta al dibujarla en un papel probabilístico normal. Se rechaza la normalidad cuando el ajuste es bajo, que corresponde a valores pequeños del estadístico del test. Dicho estadístico toma la expresión:

$$w = \frac{1}{ns^2} \left[\sum_{j=1}^h a_{j,n} (x_{(n-j+1)} - x_{(j)}) \right]^2 = \frac{A^2}{ns^2}$$

donde $ns^2 = \sum (x_i - \bar{x})^2$, h es $n/2$ si n es par y $(n-1)/2$ si n es impar. Los coeficientes $a_{j,n}$ están tabulados y $x_{(j)}$ es el valor ordenado en la muestra que ocupa el lugar j . La distribución de w está tabulada, y se rechaza la normalidad cuando su valor calculado a partir de la muestra es menor que el correspondiente valor crítico dado en las tablas. De todas formas, puede utilizarse el criterio del p-valor, rechazando la hipótesis nula de normalidad de los datos al nivel α cuando el p-valor es menor que α , y aceptándola en caso contrario.

Contrastes de normalidad de asimetría, curtosis y Jarque-Bera

Estos contrastes se basan en los coeficientes de asimetría y curtosis muestrales. Si la hipótesis de normalidad es cierta, el estadístico del contraste, que es el coeficiente de asimetría muestral $\alpha_1 = m_3 / m_2^{3/2}$, tiene una distribución asintóticamente normal de media cero y varianza $6/n$, siendo m_2 y m_3 los momentos muestrales centrados en la media de órdenes 2 y 3 respectivamente. Tenemos:

$$\alpha_1 = \frac{m_3}{m_2^{3/2}} \rightarrow N\left(0, \sqrt{\frac{6}{n}}\right)$$

Este estadístico α_1 permite contrastar la hipótesis de que los datos provienen de una distribución con simetría normal (asimetría = 0) y se basa en que si la hipótesis de normalidad es cierta, el coeficiente de asimetría estima un parámetro de la población que es cero (el coeficiente de asimetría de una distribución normal es cero). Para realizar el contraste se halla el valor k tal que $P(\alpha_1 \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico α_1 para los datos dados de la muestra es mayor que k se rechaza la hipótesis nula de simetría, y por supuesto la de normalidad.

De la misma forma, si la hipótesis de normalidad es cierta, el estadístico del contraste, que es el coeficiente de curtosis muestral $\alpha_2 = m_4 / m_2^2 - 3$, tiene una distribución asintóticamente normal de media cero y varianza $24/n$, siendo m_2 y m_4 los momentos muestrales centrados en la media de órdenes 2 y 4 respectivamente.

$$\alpha_2 = \frac{m_4}{m_2^2} - 3 \rightarrow N\left(0, \sqrt{\frac{24}{n}}\right)$$

Este estadístico α_2 permite contrastar la hipótesis de que los datos provienen de una distribución con curtosis normal (curtosis = 0) y se basa en que si la hipótesis de normalidad es cierta, el coeficiente de curtosis estima un parámetro de la población que es cero (el coeficiente de curtosis de una distribución normal es cero). Para realizar el contraste se halla el valor k tal que $P(\alpha_2 \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico α_2 para los datos dados de la muestra es mayor que k se rechaza la hipótesis nula de curtosis cero, y por supuesto la de normalidad.

Para muestras grandes, el contraste de Jarque-Bera usa los dos estadísticos anteriores mediante la consideración del estadístico de Bowman-Shelton siguiente:

$$B = n \left[\frac{\alpha_1^2}{6} + \frac{\alpha_2^2}{24} \right] \rightarrow \chi^2_2$$

Es posible utilizar para estos contrastes (como siempre) el criterio del p-valor, rechazando la hipótesis nula de normalidad de los datos al nivel α cuando el p-valor es menor que α en alguno de ellos, y aceptándola cuando el p-valor es mayor que α en los dos.

Como criterio más suave sobre la normalidad, suele considerarse normal la población cuya muestra presenta coeficientes de asimetría y curtosis comprendidos entre -2 y 2 .

La *solución habitual para la falta de normalidad* es la transformación adecuada de las variables para conseguirla.

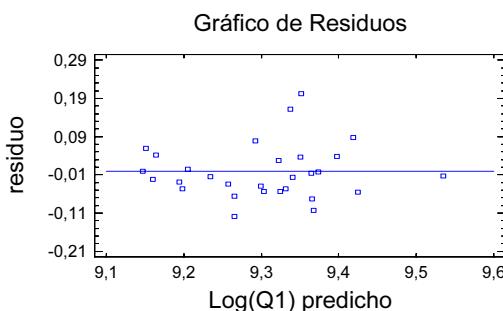
Heteroscedasticidad

En cualquier modelo multivariante suele suponerse que la variable u (término de error) es una variable aleatoria con esperanza nula ($E(u) = 0$) y matriz de covarianzas constante y diagonal ($Var(u) = \sigma^2 I_k$ matriz escalar). Es decir, que para todo t , la variable u_t tiene media cero y varianza σ^2 no dependiente de t , y además $Cov(u_i, u_j) = 0$ para todo i y para todo j distintos entre sí, pudiendo escribir $Var(u) = \sigma^2 I_k$.

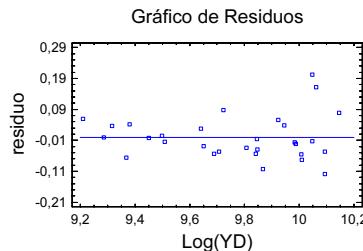
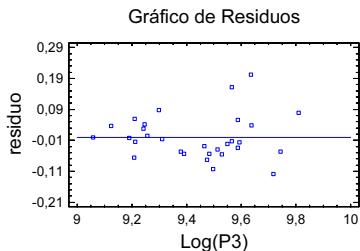
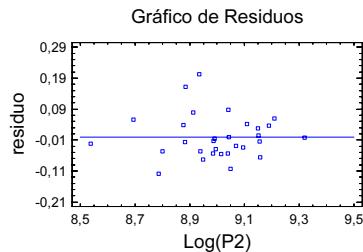
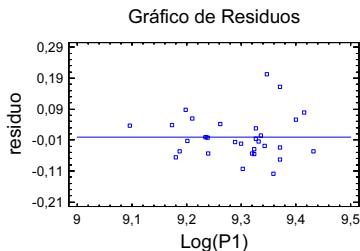
El hecho de que la varianza de u_t sea constante para todo t (que no dependa de t), se denomina hipótesis de *homoscedasticidad*. Si se relaja esta hipótesis y la varianza de u_t no es constante estamos ante la presencia de *heteroscedasticidad*. La importancia del incumplimiento de la hipótesis de homoscedasticidad radica, entre otras cosas, en que los estimadores obtenidos por MCO no son de varianza mínima aunque sigan siendo insesgados. Además, para cada variable del modelo se estimará una varianza del error.

Para analizar la heteroscedasticidad de un modelo suele comenzarse por el análisis gráfico de los residuos, siendo esenciales las gráficas de los residuos (a poder ser estudiantados) respecto de la variable endógena y respecto de las exógenas, que deben de presentar una estructura aleatoria libre de tendencia. El gráfico de los residuos contra cada variable exógena permite detectar como *variable más culpable de heteroscedasticidad* aquella cuyo gráfico se separa más de la aleatoriedad. También es un instrumento gráfico útil la gráfica de valores observados contra valores predichos, cuyos puntos han de ser lo más ajustados posible a la diagonal del primer cuadrante.

Como ejemplo, supongamos un modelo multivariante cuya variable dependiente es $\text{Log}(Q1)$ y cuyas variables independientes son $\text{Log}(P1)$, $\text{Log}(P2)$, $\text{Log}(P3)$ y $\text{Log}(YD)$. Una vez ajustado el modelo se realiza una análisis gráfico de heteroscedasticidad. El problema aparece al graficar los residuos contra los valores predichos, que muestra una estructura no demasiado aleatoria de sus puntos (este hecho nos lleva a sospechar la presencia de heteroscedasticidad) tal y como se observa en la Figura siguiente.



Para detectar qué variables son las responsables de la posible heteroscedasticidad realizamos los gráficos de residuos contra las cuatro variables explicativas. Se obtienen las Figuras siguientes:



Observándose que la menos aleatoria es la relativa a YD , ya que aumenta la dispersión del error al ir de izquierda a derecha y presenta un ajuste bueno a una recta paralela al eje X ; por tanto no tiene estructura aleatoria.

La variable YD es la candidata a provocar los problemas de heteroscedasticidad.

También suelen utilizarse los gráficos de dispersión obtenidos para las variables según se van variando los valores de una variable fijada. Si para los diferentes valores de la variable que se fija hay diferentes pautas de dispersión en el resto de las variables, puede existir heteroscedasticidad.

Aparte del análisis gráfico es necesario realizar contrastes formales de heteroscedasticidad, entre los que destacan Goldfeld-Quandt, Glesjer, Breush-Pagan, White, GARCH, ARCH y RESET de Ramsey. También suele utilizarse el test de Levenne, que comprueba la igual dispersión de la varianza en grupos formados por variables métricas. Concretamente se usa para evaluar si las varianzas de una única variable métrica son iguales a lo largo de cualquier cantidad de grupos que determinan sobre ella los valores de cualquier otra variable (que puede ser no métrica).

En general, para resolver el problema de heteroscedasticidad es conveniente tomar logaritmos. También pueden suprimirse las variables más culpables con justificación estadística y económica o introducir variables *dummy* adecuadas.

Para detectar la mejor forma funcional que sigue la varianza, se ajustan distintos modelos para las distintas formas funcionales y se toma como esquema de heteroscedasticidad aquella forma funcional para la que el ajuste es mejor.

Multicolinealidad

En un modelo multivariante suele suponerse como hipótesis que sus variables (sobre todo las variables exógenas) X_1, X_2, \dots, X_k son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se denomina hipótesis de independencia, y cuando no se cumple, decimos que el modelo presenta *multicolinealidad*.

La matriz de correlaciones es un instrumento que ayuda a detectar la presencia de multicolinealidad. Valores altos en esta matriz son síntoma de posible dependencia entre las variables implicadas.

Entre las soluciones más comunes para la multicolinealidad se tiene: ampliar la muestra, transformar las variables adecuadamente, suprimir algunas variables con justificación estadística y económica, sustitución de las variables explicativas por sus componentes principales más significativas (puntuaciones) o utilizar métodos específicos de ajuste como la regresión en cadena.

Autocorrelación

En cualquier modelo multivariante suele suponerse que la variable u (término de error) es una variable aleatoria con esperanza nula ($E(u)=0$) y matriz de covarianzas constante y diagonal ($Var(u)=\sigma^2 I_k$ matriz escalar). Es decir, que para todo t , la variable u_t tiene media cero y varianza σ^2 no dependiente de t , y además $Cov(u_i, u_j)=0$ para todo i y para todo j distintos entre sí, pudiendo escribir $Var(u)=\sigma^2 I_k$.

El hecho de que $Cov(u_i, u_j)=0$ para todo i distinto de j se denomina hipótesis de no autocorrelación. En este apartado estudiaremos el modelo lineal cuando esta hipótesis no se cumple, es decir, cuando existe *autocorrelación* o *correlación serial*.

Por tanto, en presencia de autocorrelación será necesario estimar los elementos de la matriz de varianzas covarianzas residual V . Esta tarea suele simplificarse suponiendo que las perturbaciones aleatorias del modelo siguen un determinado esquema de comportamiento que reduce el número de parámetros a estimar. Los esquemas más típicos son:

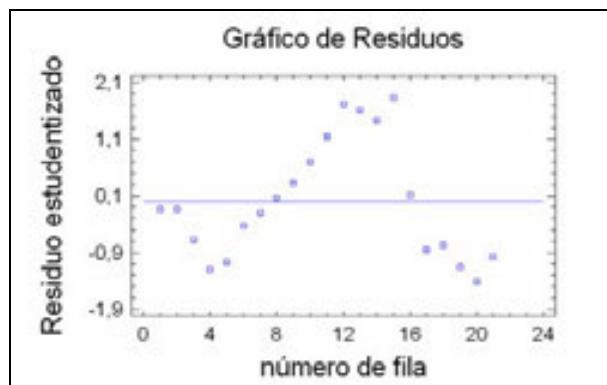
Modelo autorregresivo de orden 1 AR(1) $\rightarrow u_t = \rho u_{t-1} + e_t$

Modelo autorregresivo de orden 2 AR(2) $\rightarrow u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$

Modelo de medias móviles de orden 1 MA(1) $\rightarrow u_t = e_t + \rho e_{t-1}$

En general, las perturbaciones aleatorias pueden seguir modelos autorregresivos de medias móviles de cualquier orden, pero en el trabajo aplicado suele ser el modelo AR(1).

Para analizar la autocorrelación de un modelo suele comenzarse por el análisis gráfico de los residuos, siendo esencial la gráfica de los residuos (a poder ser estandarizados) respecto del índice tiempo (o número de fila), que debe de presentar una estructura aleatoria libre de tendencia. A continuación se presenta un caso en que existe autocorrelación (clara tendencia en el gráfico de residuos frente a nº de fila).



A parte del análisis gráfico es necesario realizar contrastes formales de autocorrelación, entre los que destacan Durbin Watson, Wallis, h-Durbin, Breusch-Godfrey y Cochrane-Orcutt.

La presencia de autocorrelación en un modelo suele solventarse mediante el método de Cochrane-Orcutt o mediante la introducción de variables *dummy* adecuadas en el modelo. Existen otros métodos menos utilizados como el método de estimación de Durbin y el procedimiento de Prais-Winsten.

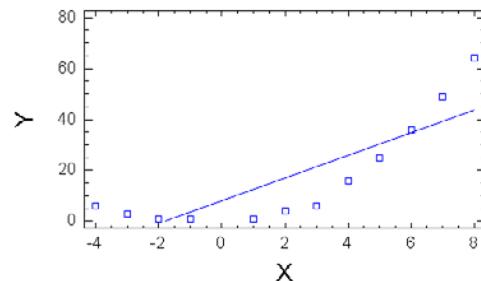
Linealidad

La linealidad es un supuesto implícito en todas las técnicas multivariantes basadas en medidas de correlación (regresión múltiple, regresión logística, análisis factorial, etc.). Como los efectos no lineales nunca están representados en el valor de la correlación, su presencia tendría efectos nocivos en el modelo multivariante.

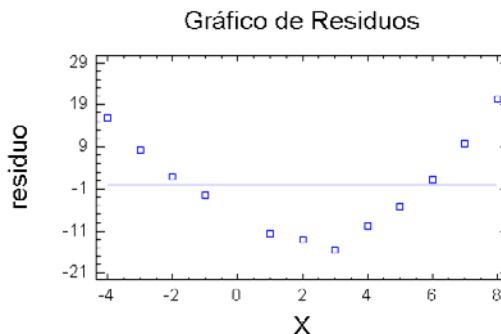
La no linealidad se resuelve tomando como modelo multivariante el modelo no lineal que se detecte que ajusta mejor las variables en estudio. El análisis gráfico permite detectar qué tipo de no linealidad puede estar presente en nuestros datos.

Los gráficos de dispersión de las variables con secuencias no lineales y los gráficos residuales con falta de aleatoriedad permiten detectar la falta de linealidad, simplemente observando su forma. Si aparecen secuencias no lineales de puntos en los gráficos de dispersión, tendremos problemas de falta de linealidad. Lo mismo ocurre si aparecen secuencias no aleatorias en los gráficos residuales.

A continuación se presenta un ejemplo de detección gráfica de falta de linealidad en un modelo variable dependiente Y e independiente X . Se comienza realizando el gráfico de dispersión de X e Y .



En la gráfica se observa que se ajusta mejor una parábola que una recta a la nube de puntos. Por otra parte, si representamos los residuos contra los valores de la variable independiente obtenemos también tendencia cuadrática tal como indica el Gráfico siguiente:



De las Figuras anteriores se deduce que hemos cometido un error de especificación en el modelo, siendo más adecuado el modelo cuadrático que el lineal. Por tanto hemos detectado la falta de linealidad y a la vez se propone el modelo que soluciona el problema.

Análisis de los residuos

Una vez construido el modelo multivariante, tendremos que contrastar entre otras las hipótesis de linealidad, normalidad, homoscedasticidad, no autocorrelación, no multicolinealidad e independencia. Los residuos van a ofrecer información relevante sobre el cumplimiento de estas hipótesis. En lo que sigue se considera Y como variable dependiente X como cualquier variable exógena.

Si el histograma de frecuencias de los residuos no se ajusta al de una normal, pueden existir valores atípicos. Eliminando los pares (X_i, Y_i) que producen los valores atípicos, se puede conseguir *normalidad* en los residuos.

Si graficamos los valores de t contra los valores de \hat{u}_t (o sea, si hacemos la gráfica cuyos puntos son los pares (t, \hat{u}_t)) y detectamos una tendencia creciente o decreciente en el grafo, puede existir *autocorrelación o correlación serial*.

Si graficamos los valores de \hat{Y}_t contra los valores de \hat{u}_t , o sea, si hacemos la gráfica cuyos puntos son los pares (\hat{Y}_t, \hat{u}_t) y detectamos una tendencia de cualquier tipo en el grafo, puede existir autocorrelación, ya que habrá correlación entre los residuos. También puede haber en este caso *heteroscedasticidad*, o también falta de *linealidad*.

Si graficamos los valores de \hat{Y}_t contra los valores de \hat{u}_t^2 , o sea, si se hace la gráfica cuyos puntos son los pares (\hat{Y}_t, \hat{u}_t^2) y detectamos una tendencia de cualquier tipo en el grafo, puede existir heteroscedasticidad.

Si graficamos los valores de X_t contra los valores de \hat{u}_t , o sea, si se hace la gráfica cuyos puntos son los pares (X_t, \hat{u}_t) y detectamos una tendencia creciente o decreciente en el grafo, puede existir autocorrelación, ya que los residuos no estarán incorrelados con las variables explicativas. También puede haber heteroscedasticidad, o falta de linealidad.

Si graficamos los valores de X_t contra los valores de \hat{u}_t^2 (o sea, si se hace la gráfica cuyos puntos son los pares (X_t, \hat{u}_t^2)) y detectamos cualquier tendencia en el grafo, puede existir heteroscedasticidad o falta de linealidad (habrá relación entre la varianza del término del error y las variables explicativas).

Estos análisis pueden realizarse también utilizando residuos estandarizados o residuos estudiantizados, que suelen ser más efectivos para detectar deficiencias en el modelo.

Los residuos estudiantizados, cuya distribución es una t de Student con $T-k-2$ grados de libertad, se usan también para detectar valores atípicos en los residuos (análisis de la normalidad de los residuos o de la mala especificación del modelo).

SPSS Y EL ANÁLISIS EXPLORATORIO DE DATOS. DATOS ATÍPICOS Y AUSENTES

ANÁLISIS EXPLORATORIO DE LOS DATOS CON SPSS. EL PROCEDIMIENTO EXPLORAR

El procedimiento *Explorar* genera estadísticos de resumen y representaciones gráficas, bien para todos los casos o bien de forma separada para grupos de casos. Existen numerosas razones para utilizar este procedimiento, por ejemplo: para inspeccionar los datos, identificar valores atípicos, obtener descripciones, comprobar supuestos y caracterizar diferencias entre subpoblaciones (grupos de casos). La inspección de los datos puede mostrar que existen valores inusuales, valores extremos, discontinuidades en los datos u otras peculiaridades. La exploración de los datos puede ayudar a determinar si son adecuadas las técnicas estadísticas que está teniendo en consideración para el análisis de los datos. La exploración puede indicar que necesita transformar los datos si la técnica necesita una distribución normal. O bien, el usuario puede decidir que necesita utilizar pruebas no paramétricas.

En cuanto a estadísticos y gráficos, se obtiene media, mediana, media recortada al 5%, error típico, varianza, desviación típica, mínimo, máximo, amplitud, amplitud intercuartil, asimetría y curtosis y sus errores típicos, intervalo de confianza para la media (y el nivel de confianza especificado), percentiles, estimador-M de Huber, estimador en onda de Andrews, estimador-M redescendente de Hampel, estimador biponderado de Tukey, los cinco valores mayores y los cinco menores, estadístico de Kolmogorov-Smirnov con el nivel de significación de Lilliefors para contrastar la normalidad y estadístico de Shapiro-Wilk. Diagramas de caja, gráficos de tallo y hojas, histogramas, diagramas de normalidad y diagramas de dispersión por nivel con pruebas de Levene y transformaciones.

Para explorar los datos, elija en los menús *Analizar* → *Estadísticos descriptivos* → *Explorar* (Figura 3-1) y seleccione una o más variables dependientes. Si lo desea, tiene la posibilidad de seleccionar una o más variables de factor cuyos valores definirán grupos de casos, seleccionar una variable de identificación para etiquetar los casos (Figura 3-2), pulsar en *Estadísticos* para obtener estimadores robustos y valores atípicos más percentiles (Figura 3-3), pulsar en *Gráficos* para obtener histogramas, pruebas y gráficos de probabilidad normal y diagramas de dispersión por nivel con estadísticos de Levene (Figura 3-4) o pulsar en *Opciones* para manipular los valores ausentes (Figura 3-5). Analizaremos el salario actual (*salario*) y los meses desde el contrato (*tiempemp*) según categoría laboral (*catlab*) etiquetando los casos según nivel educativo (*educ*) en el fichero EMPLEADOS. Pulsando *Continuar* en cada Figura, se aceptan sus especificaciones y al pulsar *Aceptar* en la Figura 3-4, se obtiene la salida del procedimiento (Figuras 3-6 a 3-17).

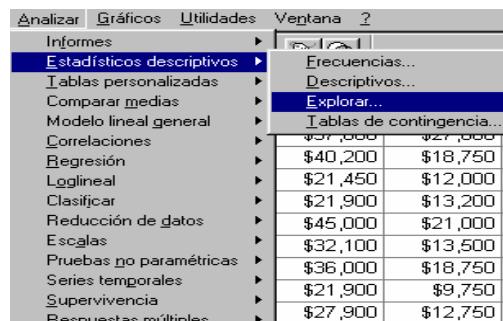


Figura 3-1



Figura 3-2

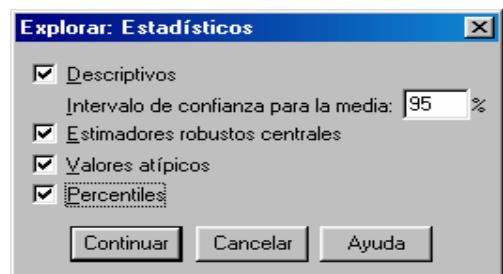


Figura 3-3

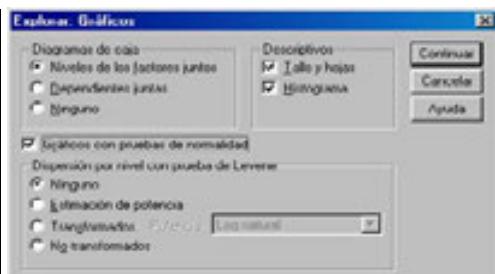


Figura 3-4

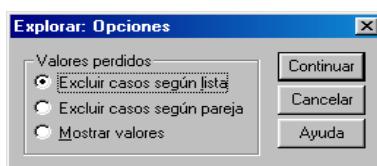


Figura 3-5

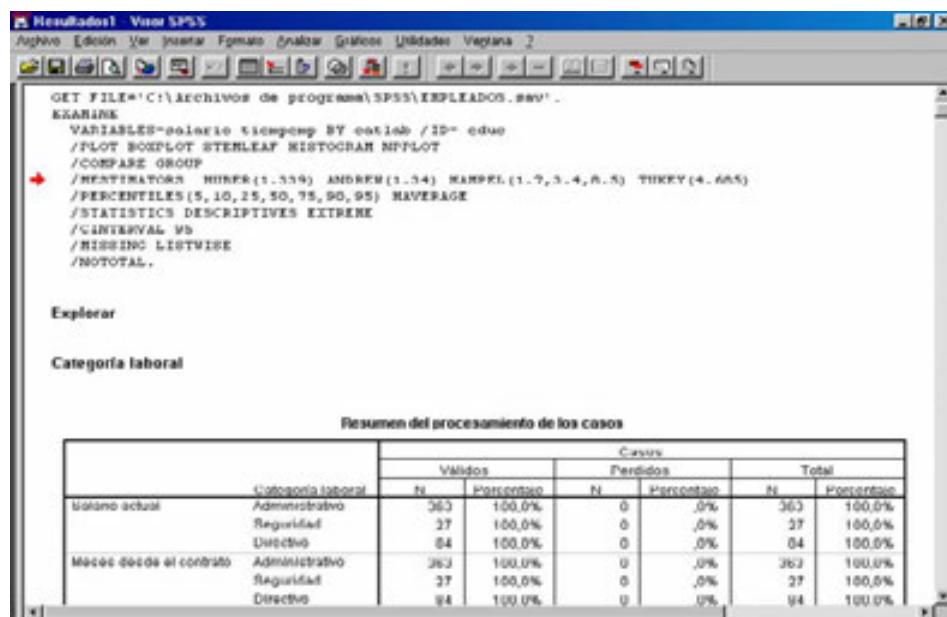


Figura 3-6

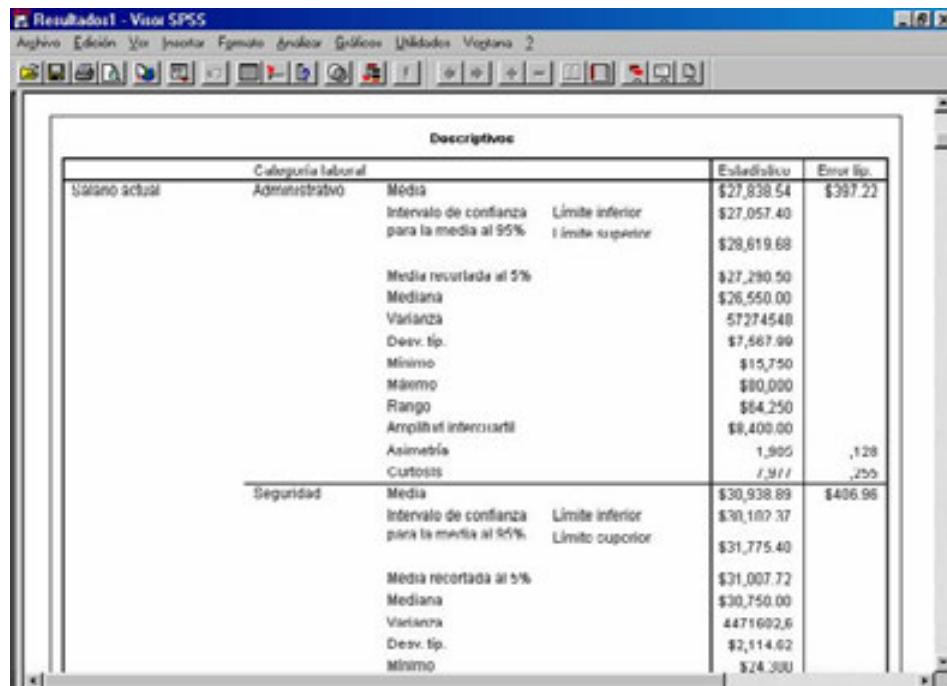


Figura 3-7

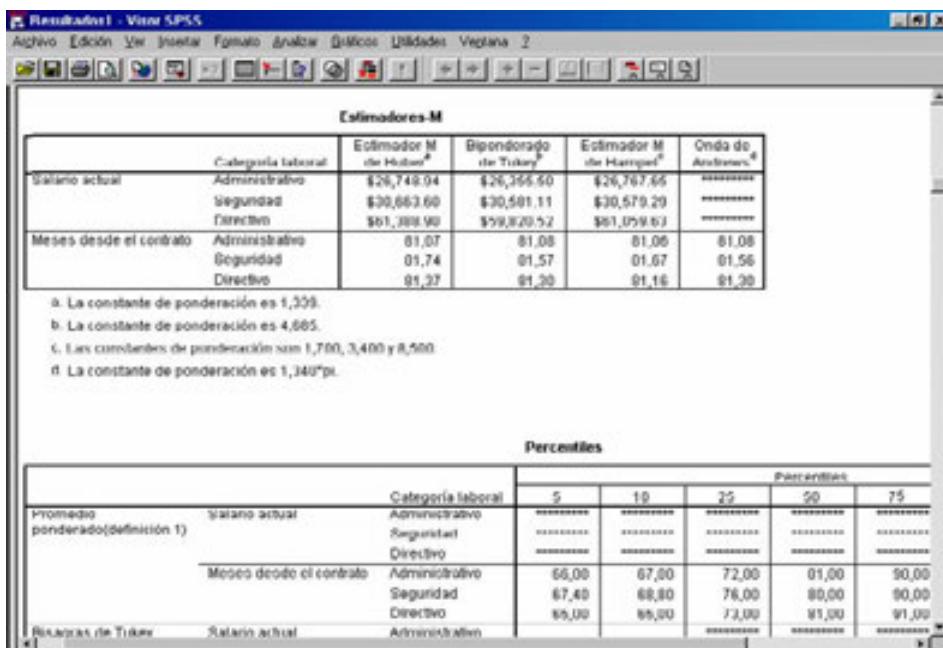


Figura 3-8

	Categoría laboral	Pruebas de normalidad			Shapiro-Wilk		
		Kolmogorov-Smirnov ^a	gl	Sig.	Estadístico	gl	Sig.
Salario actual	Administrativo	,107	363	,000			
	Seguridad	,276	27	,000	,821	27	,010**
	Directivo	,109	84	,016			
Meses dentro el contrato	Administrativo	,084	363	,000			
	Seguridad	,136	27	,200*	,945	27	,234
	Directivo	,108	84	,017			

**. Este es un límite superior de la significación verdadera.
*. Este es un límite inferior de la significación verdadera.
a. Corrección de la significación de Lilliefors

Figura 3-9

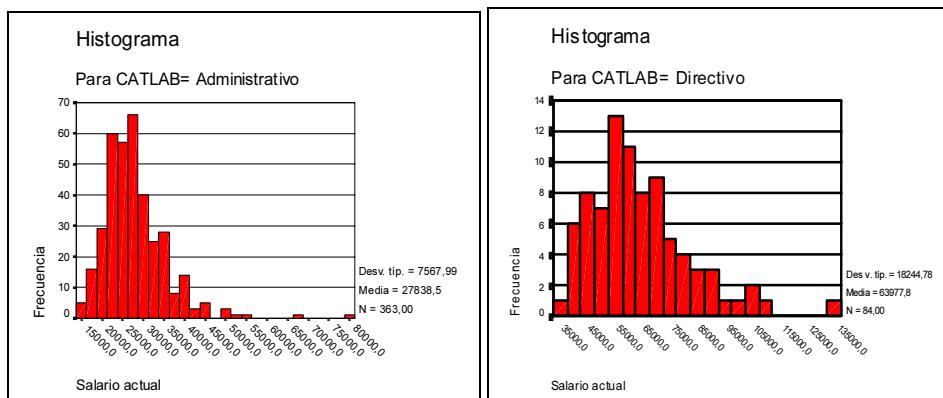


Figura 3-10

Figura 3-11

Gráficos de tallo y hojas

Salario actual Stem-and-Leaf Plot for
CATLAB= Administrativo

Frequency	Stem & Leaf
2,00	1 . 5
16,00	1 . 66666777
15,00	1 . 8899999
35,00	2 . 0000001111111111
44,00	2 . 222222222222333333
53,00	2 . 444444444444445555555555
55,00	2 . 66666666666667777777777777
35,00	2 . 8888888999999999
30,00	3 . 00000001111111
19,00	3 . 222333333
17,00	3 . 44445555
11,00	3 . 66677
8,00	3 . 8899
8,00	4 . 0000
3,00	4 . 26
12,00 Extremes	(>=43950)

Stem width: 10000
Each leaf: 2 case(s)

Figura 3-12

**Salario actual Stem-and-Leaf Plot for
CATLAB= Directivo**

Frequency	Stem & Leaf
3,00	3 . 478
15,00	4 . 001233355667788
21,00	5 . 0112344455555666788899
21,00	6 . 000011125556666788889
11,00	7 . 00023355888
4,00	8 . 1236
4,00	9 . 0127
1,00	10 . 0
4,00 Extremes	(>=103500)

Stem width: 10000
Each leaf: 1 case(s)

Figura 3-13

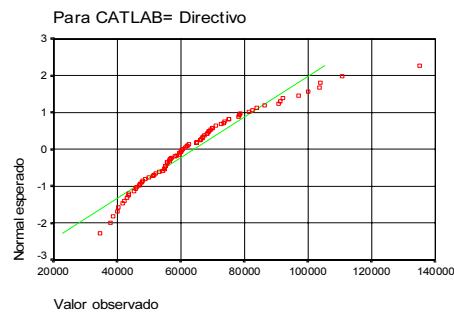
Gráfico Q-Q normal de Salario actual

Figura 3-14

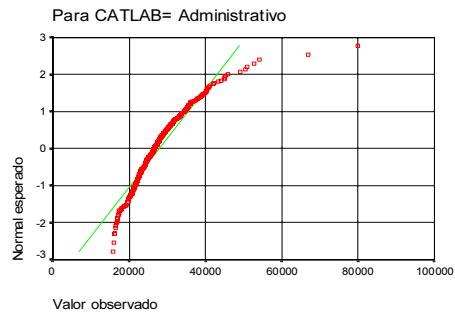
Gráfico Q-Q normal de Salario actual

Figura 3-15

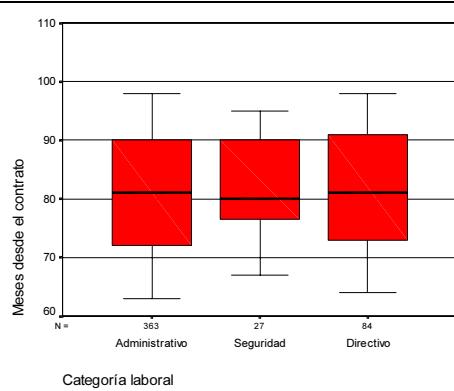


Figura 3-16

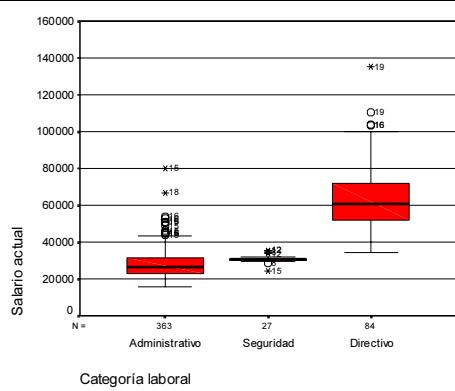


Figura 3-17

GRÁFICOS DE ANÁLISIS EXPLORATORIO CON SPSS.

Para crear un gráfico en SPSS, seleccione *Gráficos* en la barra de menús, elija el tipo de gráfico que desee en el menú *Gráficos* (Figura 3-18).

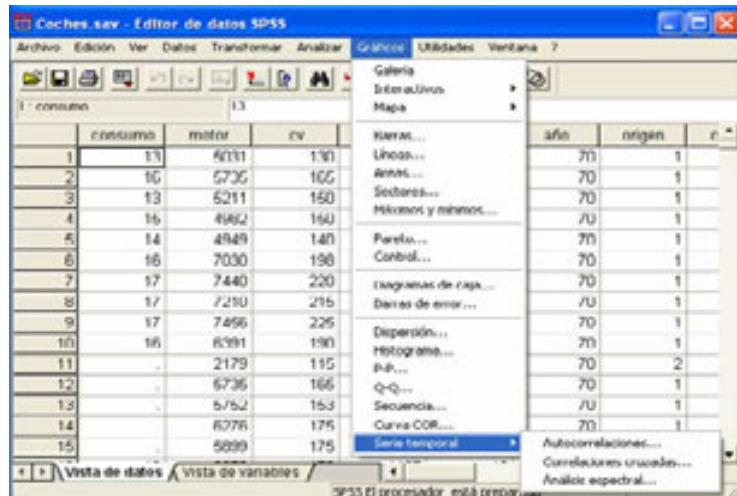


Figura 3-18

Tipos de gráficos

La Figura 3-19, que se obtiene mediante *Gráficos* → *Galería*, resume los tipos de gráficos en SPSS. Haciendo clic sobre cada gráfico se accede a su ayuda.

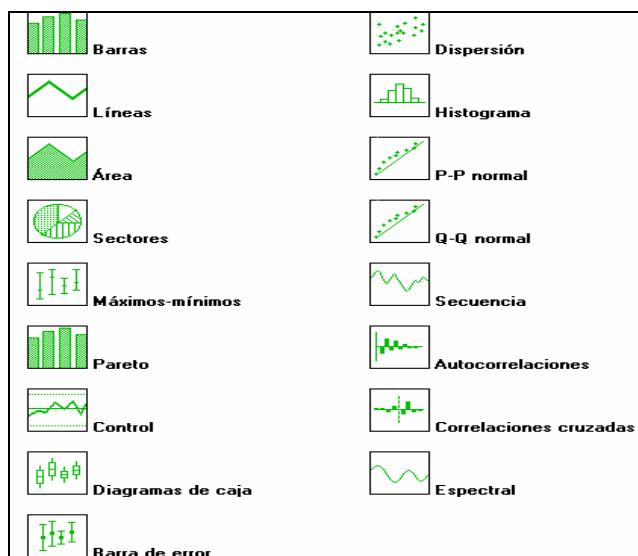


Figura 3-19

Histogramas

Seleccionando *Gráficos* → *Histograma* (Figura 3-20), se obtiene el cuadro de diálogo *Histograma* de la Figura 3-22, cuyo cuadro *Variable* permite introducir el nombre de la variable para la que se realizará el histograma (variable *Consumo* del fichero *Coches.sav*) y cuyo botón *Mostrar curva normal* permite situar sobre el histograma la campana de Gauss que mejor lo ajusta. Al hacer clic en *Aceptar* se obtiene el histograma con curva normal de la Figura 3-21.

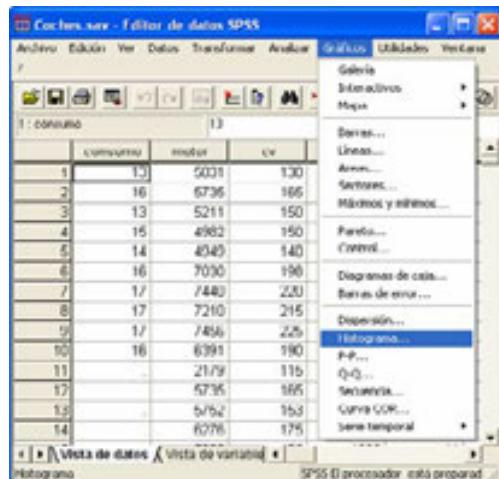


Figura 3-20

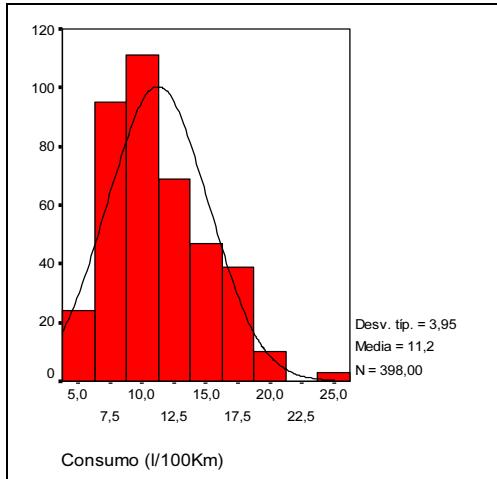


Figura 3-21

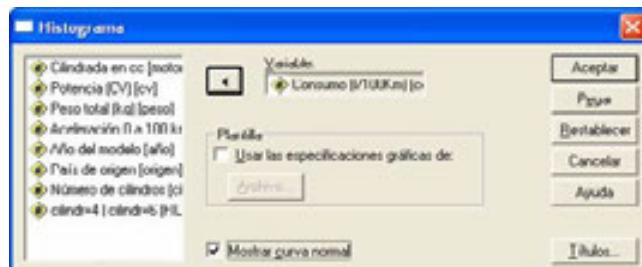


Figura 3-22

Gráficos de normalidad

Los **gráficos de probabilidad P-P** se suelen utilizar para determinar si la distribución de una variable coincide con otra distribución especificada. Si la variable seleccionada coincide con la distribución en estudio, los puntos se agruparán en torno a una línea recta. Seleccionando *Gráficos* → *P-P...* (Figura 3-23), se obtiene el cuadro de diálogo *Gráficos P-P* de la Figura 3-25, cuyo cuadro *Variable* permite introducir el nombre de la variable para la que se comprobará el ajuste a la distribución elegida (variable *Consumo* del fichero *Coches.sav*).

En el cuadro *Distribución de contraste* se elige la distribución a la que vamos a ajustar nuestros datos. Entre las distribuciones de contraste disponibles se incluyen Beta, Chi-Cuadrado, Exponencial, Gamma, Semi-Normal, Laplace, Logística, Lognormal, Normal, Pareto, t de Student, Weibull y Uniforme. Según la distribución seleccionada, podrá especificar los grados de libertad y otros parámetros en el cuadro *Parámetros de la distribución* y en las opciones *Posición* y *Escala*. La casilla *Estimar los datos* permite hallar los parámetros de la distribución de ajuste a partir de los propios datos de la variable a ajustar. En el cuadro *Transformar* es posible obtener gráficos de probabilidad para los valores transformados. Entre las opciones de transformación se incluyen *Transformación log natural*, *Tipificar los valores*, *Diferenciar* y *Diferenciar ciclo*. En los cuadros *Fórmula de estimación de la proporción* y *Rango asignado a los empates* se pueden especificar los métodos para calcular las distribuciones esperadas y para deshacer los "empates" entre múltiples observaciones con el mismo valor, tomando su media el mayor valor, el menor valor o rompiéndolos aleatoriamente. Al hacer clic en *Aceptar* se obtiene el gráfico de la Figura 3-24 que muestra un ajuste correcto de los datos a una normal.



Figura 3-23

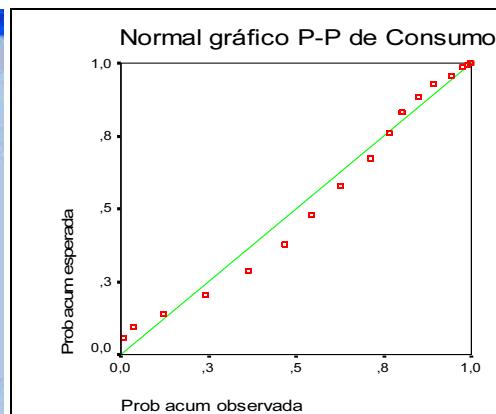


Figura 3-24

The dialog box has several sections:

- Variables:** Consumption (lit/100Km) is selected.
- Distribución de contraste:** Normal is selected.
- Parámetros de la distribución:** Estimar de los datos is checked.
- Transformar:** Options include Transformación log natural, Tipificar los valores, Diferenciar, and Diferenciar ciclo. Diferenciar is checked.
- Fórmula de estimación de la proporción:** de Blom is selected.
- Rango asignado a los empates:** Media is selected.

Figura 3-25

Los **gráficos de probabilidad Q-Q o gráficos de cuantiles** también se suelen utilizar para determinar si la distribución de una variable coincide con otra distribución especificada. Si la variable seleccionada coincide con la distribución en estudio, los puntos se agruparán en torno a una línea recta. Seleccionando *Gráficos* → *Q-Q...* (Figura 3-26), se obtiene el cuadro de diálogo *Gráficos Q-Q* de la Figura 3-28, cuyo cuadro *Variable* permite introducir el nombre de la variable para la que se comprobará el ajuste a la distribución elegida (variable *Consumo* del fichero *Coches.sav*). Los campos de la Figura 3-28 funcionan igual que para el gráfico P-P. Al hacer clic en *Aceptar* se obtiene el gráfico de la Figura 3-27 que muestra un ajuste correcto de los datos a una normal.

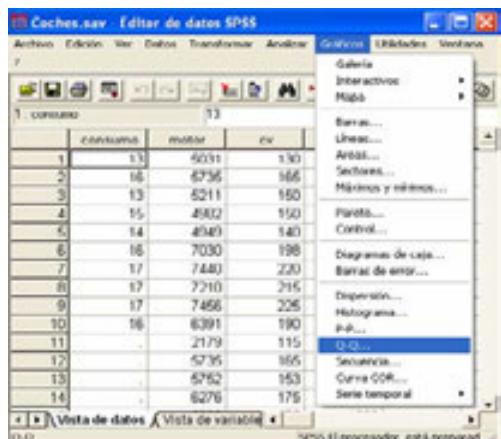


Figura 3-26

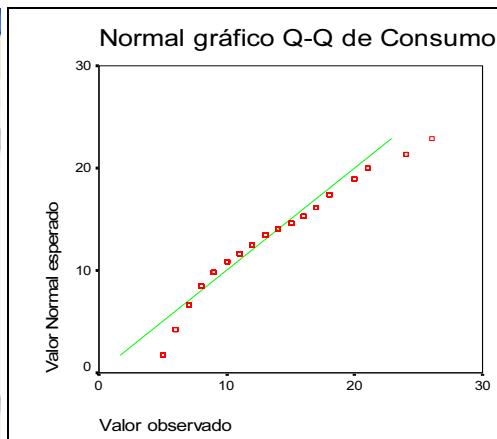


Figura 3-27

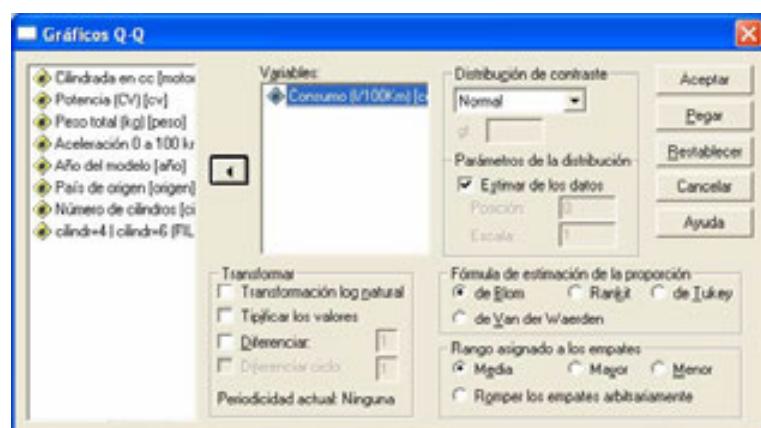


Figura 3-28

Gráficos de caja y bigotes

Vamos a representar un **gráfico de caja y bigotes simple** que resuma los datos dados por la variable potencia de los automóviles (*cv*) del fichero de datos sobre coches (COCHES), presentando sobre este gráfico la media, la mediana y los valores atípicos. Comenzamos cargando en memoria el fichero COCHES mediante *Archivo → Abrir → Coches*. A continuación seleccionamos *Gráficos → Diagramas de caja* y elegimos *Simple* y *Resúmenes para distintas variables* en la Figura 3-29. Al pulsar *Definir* se obtiene la Figura 3-30 en la que elige la variable *cv* para representar en cajas. Al pulsar *Aceptar* se obtiene la Figura 3-31.



Figura 3-29

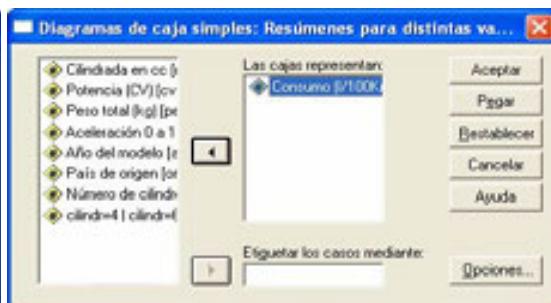


Figura 3-30

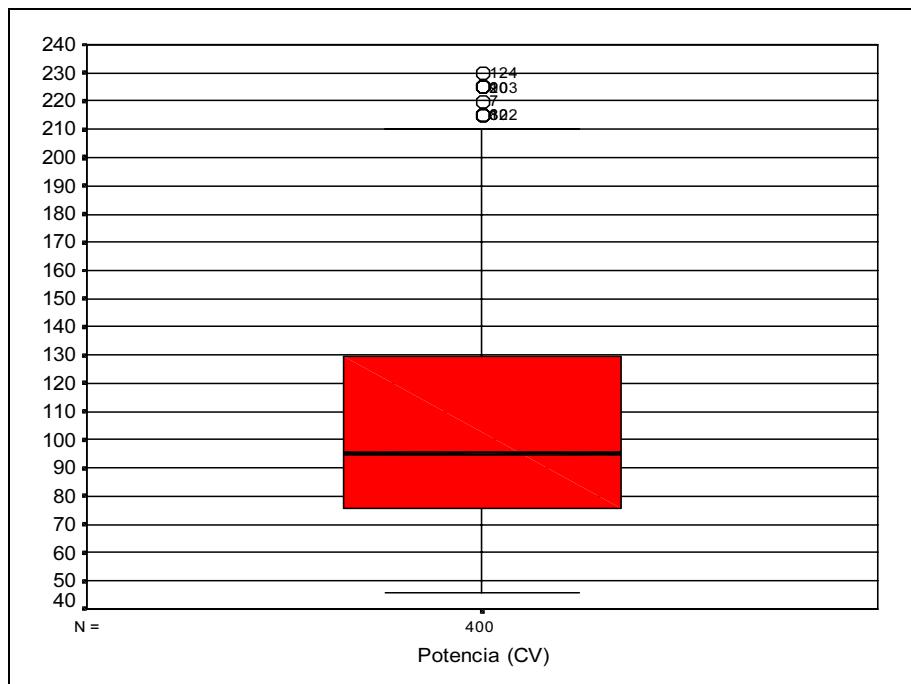


Figura 3-31

La línea horizontal del interior de la caja corresponde a la mediana, cuyo valor será cercano a 95 y tenemos cuatro valores atípicos representados por puntos situados por encima del bigote superior. El primer cuartil corresponde más o menos al valor 75 (línea horizontal inferior de la caja) y el tercero al valor 130 (línea horizontal superior de la caja). A la vista del gráfico podemos decir que la potencia de los coches varía entre 45 y 230 caballos aproximadamente, y que el 50% de los coches tiene su potencia entre 75 y 130 caballos. La cantidad que deja a su izquierda y a su derecha el mismo número de valores de la variable potencia (valor mediano) es 85. Por otra parte, hay cuatro coches cuya potencia es anormalmente grande y la distribución de la variable potencia de los coches es ligeramente asimétrica hacia la derecha (hay más trozo de caja por encima de la línea mediana).

Ahora representamos un gráfico de caja y bigotes que resume los datos dados para la variable potencia de los automóviles (*cv*) del fichero de datos sobre coches (COCHES), clasificados en tres gráficos simples de caja y bigotes. Esta clasificación vendrá dada por los valores 1, 2 y 3 de la variable *origen* (región de origen de los coches), cuyas etiquetas respectivas son: EE.UU., Europa y Japón.

Cargamos en memoria el fichero COCHES mediante *Archivo → Abrir → Coches*. A continuación seleccionamos *Gráficos → Diagramas de caja* y elegimos *Simple y Resúmenes para grupos de casos* en la Figura 3-32. Al pulsar *Definir* se obtiene la Figura 3-33 en la que se elige como variable *cv* y como eje de categorías *origen*. Al pulsar *Aceptar* se obtiene la Figura 3-34.

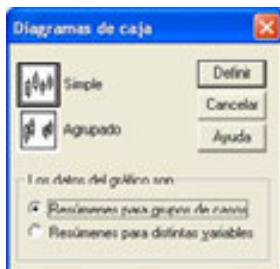


Figura 3-32

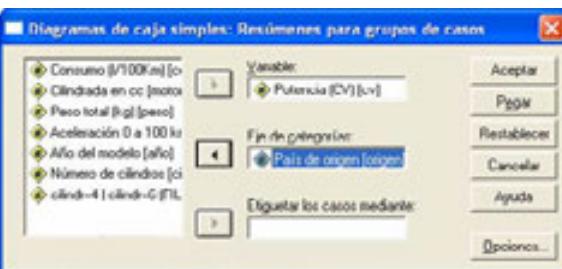


Figura 3-33

Las líneas horizontales del interior de las cajas corresponden a las medianas, cuyos valores son cercanos a 105, 78 y 73 respectivamente. Tenemos dos valores atípicos en la potencia de los coches europeos, el primer cuartil de cada grupo corresponde más o menos a los valores 87, 70 y 67 (líneas horizontales inferiores de las cajas) y el tercero de cada grupo a los valores 150, 92 y 95 (líneas horizontales superiores de las cajas). A la vista del gráfico las tres distribuciones son asimétricas a la derecha. Observando las cajas se deduce que es mayor el intervalo que hay entre la mediana y el tercer cuartil que el comprendido entre el primer cuartil y la mediana, luego habrá más variedad para elegir en cuanto a la potencia en los coches si nos situamos en la zona cuya potencia es superior a la mediana.

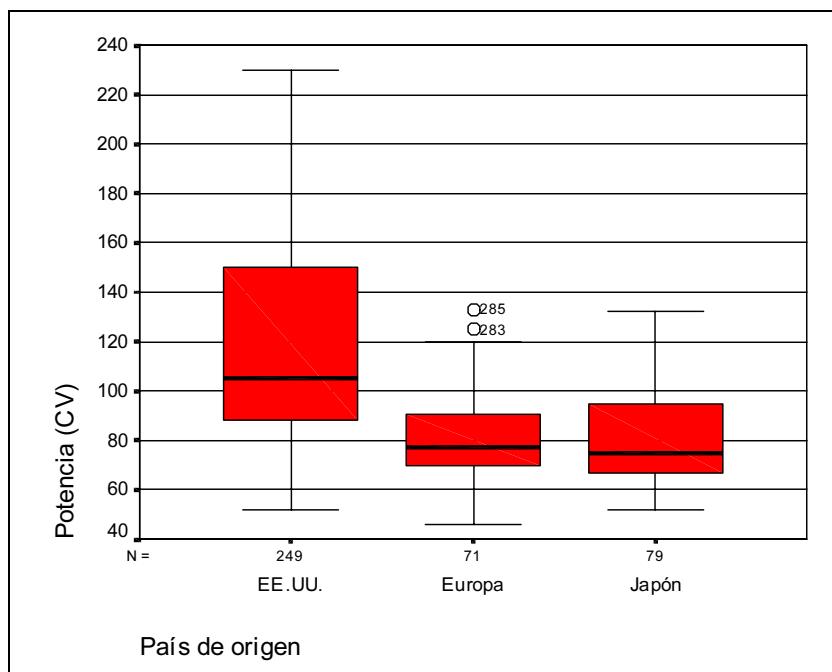


Figura 3-34

Gráficos de control para la detección de casos atípicos

SPSS permite realizar el gráfico de control *Individuos* y *Rango móvil* que se utiliza para la detección univariante de casos atípicos. Vamos a representar este tipo de gráfico para los datos dados por la variable potencia de los automóviles (*cv*) del fichero de datos sobre coches (COCHES). Comenzamos cargando en memoria el fichero COCHES mediante *Archivo* → *Abrir* → *Coches*. A continuación seleccionamos *Gráficos* → *Control* (Figura 3-35) y elegimos *Individuos, rango móvil* situando en *Organización de los datos* la opción *Los datos son unidades* (Figura 3-36).

Al pulsar *Definir* se obtiene la Figura 3-37 en la que elige la variable *cv* para representar en cajas. El botón *Estadísticos* (Figura 3-38) permite definir los límites de especificación para el gráfico y seleccionar diversos índices de funcionalidad y de rendimiento (normalmente se toman las opciones por defecto). La casilla *Amplitud* permite elegir el número de casos utilizado para calcular el rango móvil. Por ejemplo, si la amplitud (la duración) es tres, se utilizan el caso actual y los dos casos previos. El valor de la amplitud (la duración) también se utiliza para calcular los límites de control de ambos gráficos. Con el botón *Opciones* (Figura 3-39) es posible especificar el número de desviaciones típicas (sigmas) utilizadas para calcular los límites de control y añadir al gráfico límites de control fijos. Al pulsar *Aceptar* se obtiene la Figura 3-40, que muestra varios puntos fuera de control (valores atípicos), sobre todo en el extremo superior de la variable.



Figura 3-35



Figura 3-36

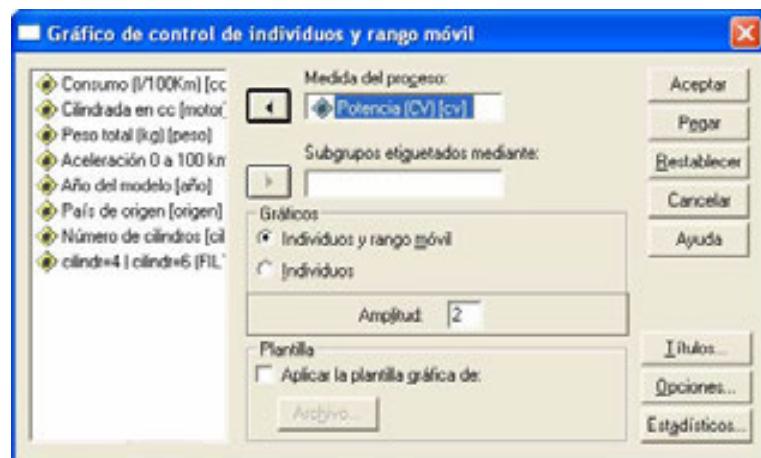


Figura 3-37

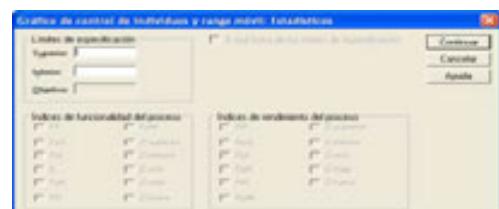


Figura 3-38

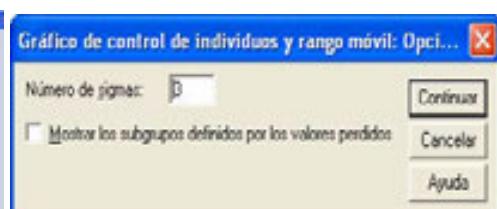


Figura 3-39

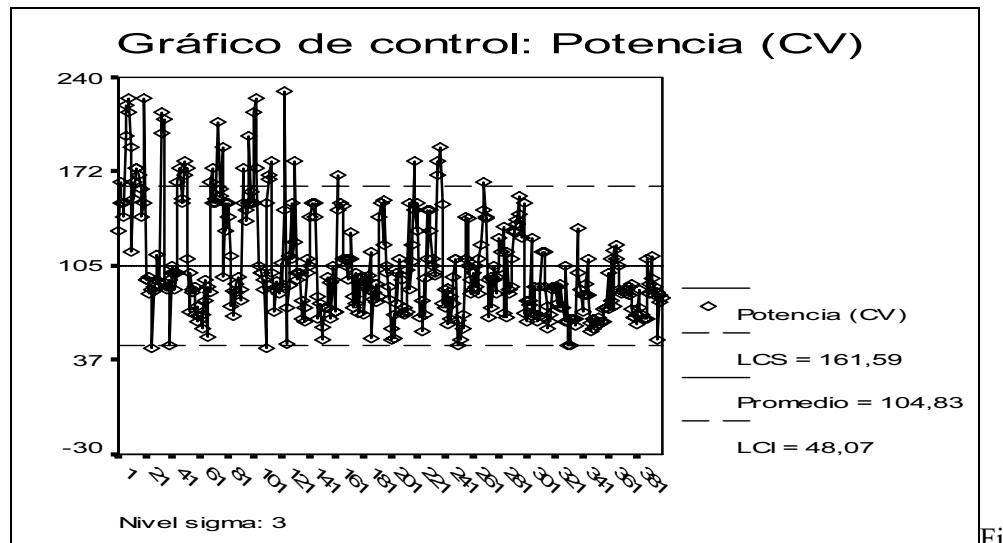


Figura 3-40

Gráficos de dispersión

Este tipo de gráfico es el más utilizado para ver la relación entre dos o más variables. Vamos a representar este tipo de gráfico para los datos dados por la variable potencia de los automóviles (*cv*) según su *Consumo* del fichero de datos sobre coches (COCHES). Comenzamos cargando en memoria el fichero COCHES mediante *Archivo → Abrir → Coches*. A continuación seleccionamos *Gráficos → Dispersion* (Figura 3-41) y elegimos *Simple* en la Figura 3-43. Al pulsar *Definir* se obtiene la Figura 3-42 en la que elige la variable *cv* en el eje X y la variable *Consumo* en el eje Y. El botón *Opciones* (Figura 3-44) permite el tratamiento de los valores perdidos. Al pulsar *Aceptar* se obtiene la Figura 3-45.

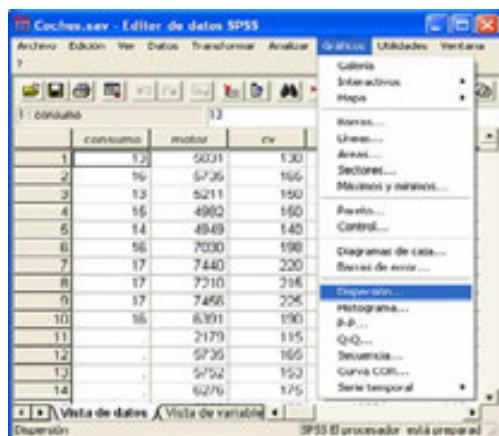


Figura 3-41

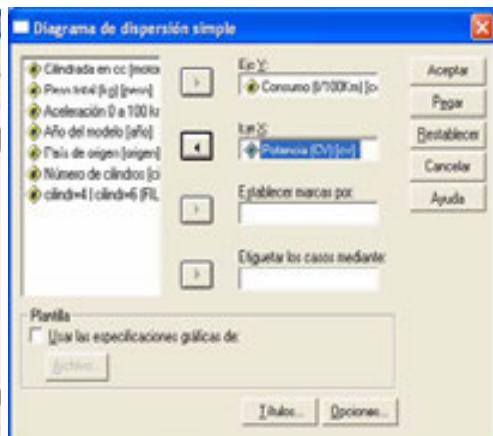


Figura 3-42

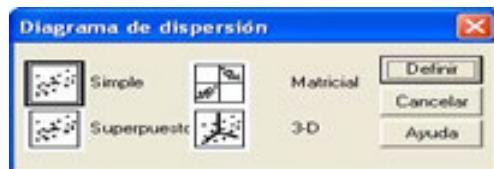


Figura 3-43

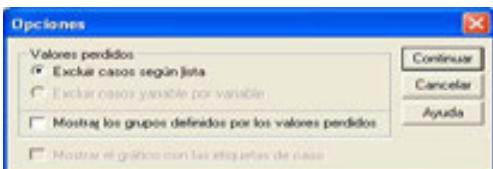


Figura 3-44

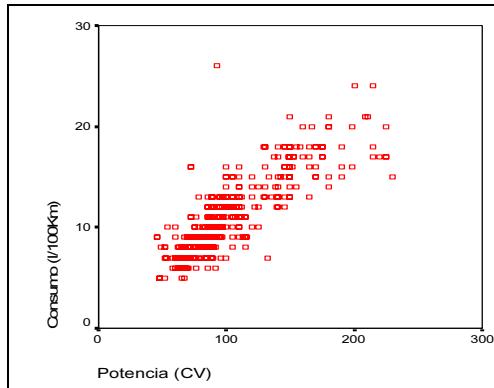


Figura 3-45

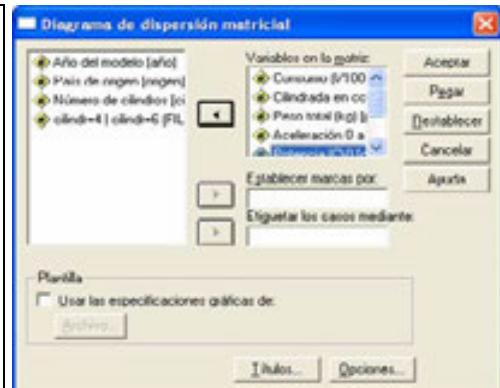


Figura 3-46

Si en la Figura 3-43 elegimos Matricial, pulsamos *Definir* y rellenamos la pantalla *Diagrama de dispersión matricial* como se indica en la Figura 3-46, al pulsar *Aceptar* se obtiene el gráfico de dispersión matricial de la Figura 3-47, que muestra la relación entre todas las variables (consumo, cilindrada, peso, aceleración y potencia).

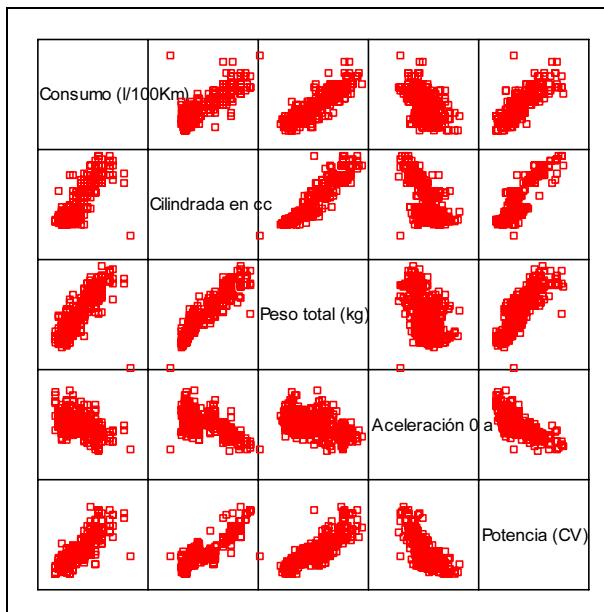


Figura 3-47

TRATAMIENTO DE LOS VALORES AUSENTES CON SPSS

Ya sabemos que tras observar la presencia de datos ausentes en una distribución será necesario detectar si estos se distribuyen aleatoriamente. Será necesario detectar que el efecto de los datos ausentes es importante mediante pruebas formales de aleatoriedad o mediante la técnica de la matriz de correlaciones dicotomizada.

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico con ellos. Habrá que decidir entre la imputación de la información faltante y su supresión ordenada.

Diagnóstico de los datos ausentes con SPSS. El procedimiento Prueba T para muestras independientes

Una primera prueba para valorar los datos ausentes para una única variable Y consiste en formar dos grupos de valores para Y, los que tienen datos ausentes y los que no los tienen. A continuación, para cada variable X distinta de Y, se realiza un test para determinar si existen diferencias significativas entre los dos grupos de valores determinados por la variable Y (ausentes y no ausentes) sobre X. Si vamos considerando como Y cada una de las variables del análisis y repitiendo el proceso anterior se encuentra que todas las diferencias son no significativas, se puede concluir que los datos ausentes obedecen a un *proceso completamente aleatorio* y por tanto pueden realizarse análisis estadísticos fiables con nuestras variables *imputando los datos ausentes* por los métodos que se verán más adelante.

El procedimiento *Prueba T* para muestras independientes compara las medias de dos grupos de casos. Para esta prueba, idealmente los sujetos deben asignarse aleatoriamente a dos grupos, de forma que cualquier diferencia en la respuesta sea debida al tratamiento (o falta de tratamiento) y no a otros factores. Este caso no ocurre si se comparan los ingresos medios para hombres y mujeres. El sexo de una persona no se asigna aleatoriamente. En estas situaciones, debe asegurarse de que las diferencias en otros factores no enmascaren o resalten una diferencia significativa entre las medias. Las diferencias de ingresos medios pueden estar sometidas a la influencia de factores como los estudios y no solamente el sexo. En cuanto a estadísticos, para cada variable a contrastar da para cada variable su tamaño de la muestra, media, desviación típica y error típico de la media. Para la diferencia entre las medias da media, error típico e intervalo de confianza (puede especificar el nivel de confianza). En cuanto a contrastes ofrece la prueba de Levene sobre la igualdad de varianzas y pruebas T de varianzas combinadas y separadas sobre la igualdad de las medias.

En cuanto a los datos, los valores de la variable cuantitativa de interés se hallan en una única columna del archivo de datos. El procedimiento utiliza una variable de agrupación con dos valores para separar los casos en dos grupos. La variable de agrupación puede ser numérica (valores como 1 y 2, ó 6,25 y 12,5) o de cadena corta (como SÍ y NO). También puede usar una variable cuantitativa, como la EDAD, para dividir los casos en dos grupos especificando un punto de corte (el punto de corte 21 divide la EDAD en un grupo de menos de 21 años y otro de más de 21).

Para obtener una prueba T para muestras independientes, elija en los menús *Analizar* → *Comparar medias* → *Prueba T para muestras independientes* (Figura 3-48). Seleccione una o más variables de contraste cuantitativas (Figura 3-49). Se calcula una prueba T diferente para cada variable. Seleccione una sola variable de agrupación y pulse en *Definir grupos* para especificar dos códigos para los grupos que deseé comparar. Contrastaremos la igualdad de medias de la variable *educ* (nivel educativo) según *sexo* en el fichero EMPLEADOS. Si lo desea, puede pulsar en *Opciones* (Figura 3-50) para controlar el tratamiento de los datos perdidos y el nivel del intervalo de confianza. Al pulsar *Aceptar* en la Figura 3-49 se obtiene la salida con la sintaxis del procedimiento y un resumen de estadísticos para la muestra (Figura 3-51) y los resultados del contraste de la T con significatividad para la diferencia de nivel educativo por sexo en media y con un intervalo de confianza para las diferencias medias, que evidentemente no contiene el valor 0 (Figura 3-52). Existe diferencia de nivel educativo en media entre hombres y mujeres.

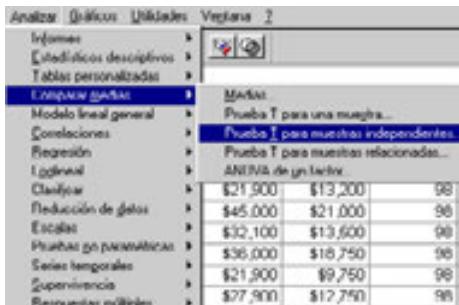


Figura 3-48



Figura 3-49

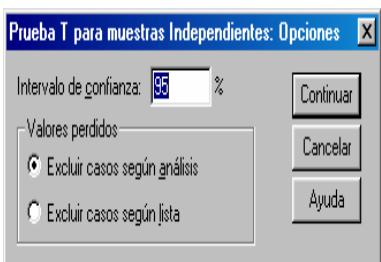


Figura 3-50

T-TEST GROUPS=sexo(1 2) /MISSING=ANALYSIS /VARIABLES=educ /CRITERIA=CIN(.95) .				
Prueba T				
Estadísticos de grupo				
Sexo	N	Media	Desviación tip.	Error típ. de la media
Nivel educativo	Hombre	258	14,43	,298 .19
	Mujer	216	12,37	,232 .16

Figura 3-51

		Prueba de muestras independientes								
		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de media	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
									Inferior	Superior
Nivel educativo	Se han asumido varianzas iguales	17,884	,000	8,276	472	,000	2,06	,25	1,57	2,55
	No se han asumido varianzas iguales			8,458	469,6	,000	2,06	,24	1,58	2,54

Figura 3-52

Diagnóstico de los datos ausentes con SPSS. El procedimiento Correlaciones bivariadas

El procedimiento *Correlaciones bivariadas* calcula el coeficiente de correlación de Pearson, la rho de Spearman y la tau-b de Kendall con sus niveles de significación. Las correlaciones miden cómo están relacionadas las variables o los órdenes de los rangos. Antes de calcular un coeficiente de correlación, inspeccione los datos para detectar valores atípicos (que pueden producir resultados equívocos) y evidencias de una relación lineal. El coeficiente de correlación de Pearson es una medida de asociación lineal. Dos variables pueden estar perfectamente relacionadas, pero si la relación no es lineal, el coeficiente de correlación de Pearson no será un estadístico adecuado para medir su asociación. En cuanto a estadísticos, para cada variable se calcula: número de casos sin valores perdidos, desviación típica y media. Para cada pareja de variables se calcula: coeficiente de correlación de Pearson, rho de Spearman, tau-b de Kendall, productos cruzados de las desviaciones y covarianzas. En cuanto a los datos, utilice variables cuantitativas simétricas para el coeficiente de correlación de Pearson y variables cuantitativas o variables con categorías ordenadas para la rho de Spearman y la tau-b de Kendall.

Para obtener correlaciones bivariadas elija en los menús *Analizar → Correlaciones → Bivariadas* (Figura 3-53) y seleccione dos o más variables numéricas (Figura 3-54). También se encuentran disponibles en la Figura 3-54 las siguientes opciones:

Coeficientes de correlación: Para las variables cuantitativas, normalmente distribuidas, seleccione el coeficiente de correlación de Pearson. Si los datos no están normalmente distribuidos o tienen categorías ordenadas, seleccione la tau-b de Kendall o de Spearman, que miden la asociación entre órdenes de rangos. Los coeficientes de correlación pueden estar entre -1 (una relación negativa perfecta) y +1 (una relación positiva perfecta). Un valor 0 indica que no existe una relación lineal. Al interpretar los resultados, se debe evitar extraer conclusiones de causa-efecto a partir de una correlación significativa.

Prueba de significación: Se pueden seleccionar las probabilidades bilaterales o las unilaterales. Si conoce de antemano la dirección de la asociación, seleccione *Unilateral*. Si no es así, seleccione *Bilateral*.

Marcar las correlaciones significativas: Los coeficientes de correlación significativos al nivel 0,05 se identifican por medio de un solo asterisco y los significativos al nivel 0,01 se identifican con dos asteriscos.

Opciones: Permite elegir estadísticos y manejar valores perdidos (Figura 3-55).

Previamente es necesario cargar en memoria el fichero de nombre MUNDO mediante *Archivo → Abrir → Datos*. Este fichero contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo entre los que se encuentra la esperanza de vida femenina (variable *espvidaf*), la esperanza de vida masculina (variable *espvidam*) y el porcentaje de alfabetización (*alfabet*).

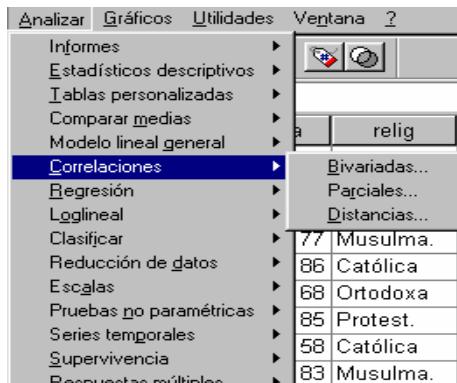


Figura 3-53

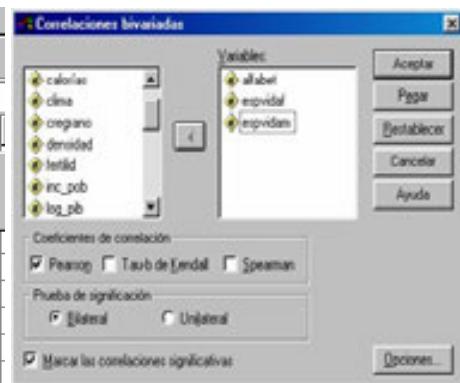


Figura 3-54

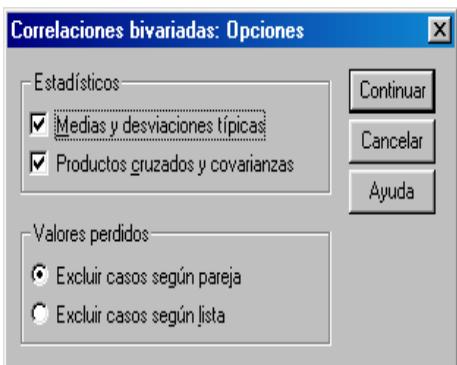


Figura 3-55



Figura 3-55

Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 3-54 para obtener en primer lugar la sintaxis del procedimiento y los estadísticos descriptivos (Figura 3-57). A continuación se obtiene la matriz de correlaciones y covarianzas con los coeficientes de significación de cada correlación (Figura 3-57).

Se observa la alta correlación entre las tres variables y la elevada significatividad de todos los coeficientes de correlación dos a dos.

Correlaciones				
		Alfabetización (%)	Esperanza de vida femenina	Esperanza de vida masculina
Alfabetización (%)	Correlación de Pearson	1,000	,865**	,809**
	Sig. (bilateral)	,	,000	,000
	Suma de cuadrados y productos cruzados	55505,888	22356,636	18345,710
	Covarianza	523,640	210,912	173,073
	N	107	107	107
Esperanza de vida femenina	Correlación de Pearson	,865**	1,000	,982**
	Sig. (bilateral)	,000	,	,000
	Suma de cuadrados y productos cruzados	22356,636	12070,349	10400,404
	Covarianza	210,912	111,762	96,300
	N	107	109	109
Esperanza de vida masculina	Correlación de Pearson	,809**	,982**	1,000
	Sig. (bilateral)	,000	,000	,
	Suma de cuadrados y productos cruzados	18345,710	10400,404	9286,257
	Covarianza	173,073	96,300	85,984
	N	107	109	109

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

Figura 3-57

Técnicas de imputación de datos ausentes con SPSS. El procedimiento Reemplazar los valores perdidos

Probada la existencia de aleatoriedad en los datos ausentes puede que decidamos imputar la información faltante. Las *observaciones perdidas* pueden causar problemas en los análisis y algunas medidas de series temporales no se pueden calcular si hay valores perdidos en la serie. *Reemplazar valores perdidos* crea nuevas variables a partir de otras existentes, reemplazando los valores perdidos por estimaciones calculadas mediante uno de los distintos métodos posibles de imputación. Los nombres de las nuevas variables por defecto se componen de los seis primeros caracteres de la variable existente utilizada para crearlas, seguidos de un carácter de subrayado y de un número secuencial. Por ejemplo, para la variable PRECIO, el nombre de la nueva variable sería PRECIO_1. Las nuevas variables conservan cualquier etiqueta de valor ya definida en las variables originales.

Para reemplazar los valores perdidos para las variables de series temporales, elija en los menús: *Transformar* → *Reemplazar valores perdidos* (Figura 3-58) y seleccione el método de estimación que deseé utilizar para reemplazar los valores perdidos (Figura 3-59). Seleccione la variable o variables para las que desea reemplazar los valores perdidos. Se pueden imputar los valores perdidos sustituyéndolos por la media de la serie, por la media o mediana de los puntos adyacentes, por interpolación lineal o por tendencia lineal en el punto.

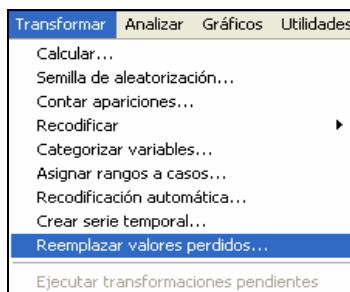


Figura 3-58

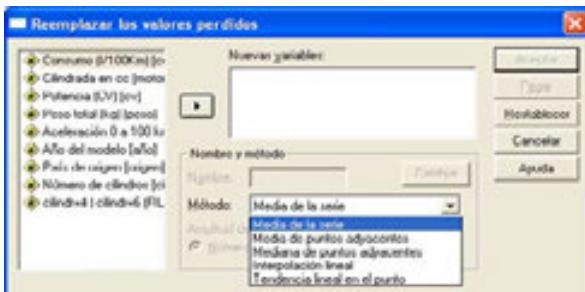


Figura 3-59

Supresión de los datos ausentes con SPSS

En el tratamiento de los datos ausentes también se puede tomar la decisión de eliminarlos. Podemos comenzar incluyendo sólo en el análisis las observaciones (casos) con datos completos (filas cuyos valores para todas las variables sean válidos), es decir, cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se denomina *aproximación de casos completos* o *supresión de casos según lista* y suele ser el método por defecto en la mayoría del software estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, porque su supresión provocaría una muestra representativa de la información total. En caso contrario se reduciría mucho el tamaño de la muestra a considerar para el análisis y no sería representativa de la información completa. Otro método consiste en la *supresión de datos según pareja*, es decir, se trabaja con todos los casos (filas) posibles que tengan valores válidos para cada par de variables que se consideren en el análisis independientemente de lo que ocurra en el resto de las variables. Este método elimina menos información y se utiliza siempre en cualquier análisis bivariante o transformable en bivariante. Otro método adicional consiste en *suprimir los casos (filas) o variables (columnas)* que peor se comportan respecto a los datos ausentes. Nuevamente es necesario sopesar la cantidad de datos a eliminar. Debe siempre considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable o conjunto de casos en el análisis estadístico. SPSS, en la mayoría de sus procedimientos ofrece un botón *Opciones* que permite utilizar el método de supresión de datos ausentes que se deseé (Figura 3-5 en el procedimiento *Explorar*, Figura 3-44 en el procedimiento *Diagrama de dispersión simple*, Figura 3-55 en el procedimiento *Correlaciones bivariadas*, etc.).

EL PROCEDIMIENTO FRECUENCIAS DE SPSS

El procedimiento *Frecuencias* proporciona estadísticos y representaciones gráficas que resultan útiles para describir muchos tipos de variables. Es un buen procedimiento para una inspección inicial de los datos. En cuanto a estadísticos y gráficos proporciona frecuencias, porcentajes, porcentajes acumulados, media, mediana, moda, suma, desviación típica, varianza, amplitud, valores mínimo y máximo, error típico de la media, asimetría y curtosis (ambos con sus errores típicos), cuartiles, percentiles especificados por el usuario, gráficos de barras, gráficos de sectores e histogramas.

Para los informes de frecuencias y los gráficos de barras, puede organizar los diferentes valores en orden ascendente o descendente u ordenar las categorías por sus frecuencias. Es posible suprimir el informe de frecuencias cuando una variable posee muchos valores diferentes. Puede etiquetar los gráficos con las frecuencias (la opción por defecto) o con los porcentajes. En cuanto a los datos, utilice códigos numéricos o cadenas cortas para codificar las variables categóricas (medidas de nivel nominal u ordinal).

Para ejecutar el procedimiento, elija en los menús *Analizar* → *Estadísticos descriptivos* → *Frecuencias* (Figura 3-60) y seleccione una o más variables categóricas o cuantitativas (Figura 3-61). Previamente es necesario cargar en memoria el fichero de nombre MUNDO mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo entre los que se encuentra la población urbana (variable *urbana*). Si lo desea, tiene la posibilidad de pulsar en *Estadísticos* para obtener estadísticos descriptivos para las variables cuantitativas (Figura 3-62), pulsar en *Gráficos* para obtener gráficos de barras, gráficos de sectores e histogramas (Figura 3-63) o pulsar en *Formato* para determinar el orden en el que se muestran los resultados (Figura 3-64).

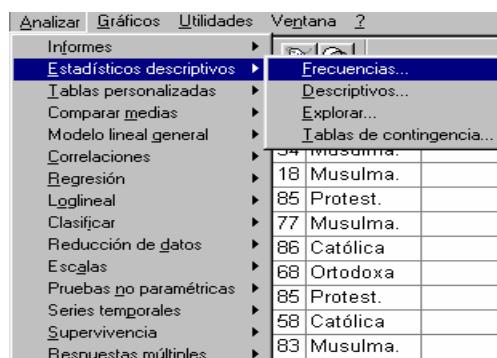


Figura 3-60

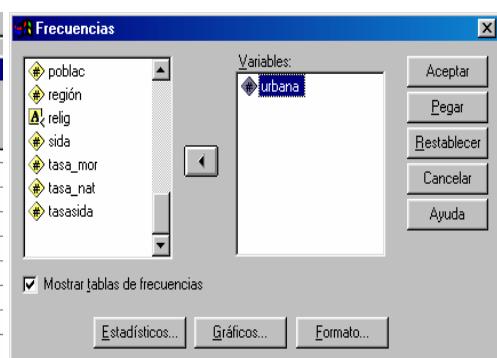


Figura 3-61



Figura 3-62

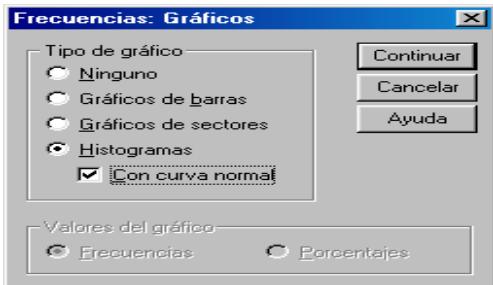


Figura 3-63

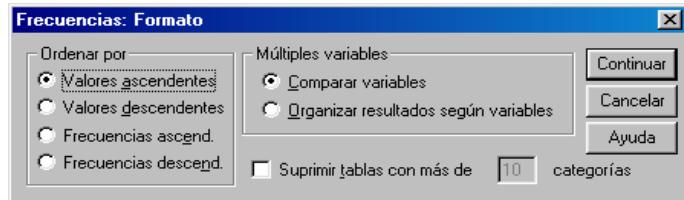


Figura 3-64

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. El botón *Continuar* permite aceptar las asignaciones, el botón *Cancelar* permite ignorarlas y el botón *Pegar* envía la sintaxis del procedimiento a la ventana de sintaxis. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 3-61 para obtener los resultados del procedimiento según se muestra en la Figura 3-65. En la parte izquierda del Visor podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. La Figura 3-66 muestra el histograma de frecuencias con curva normal.

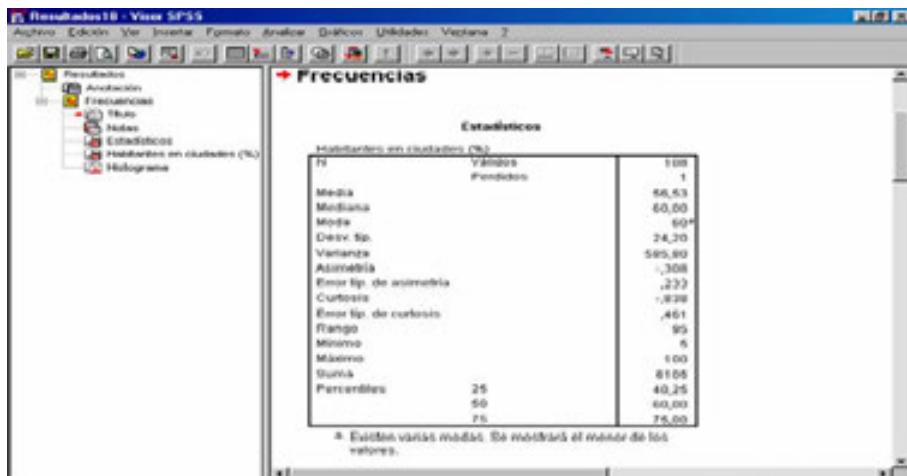


Figura 3-65

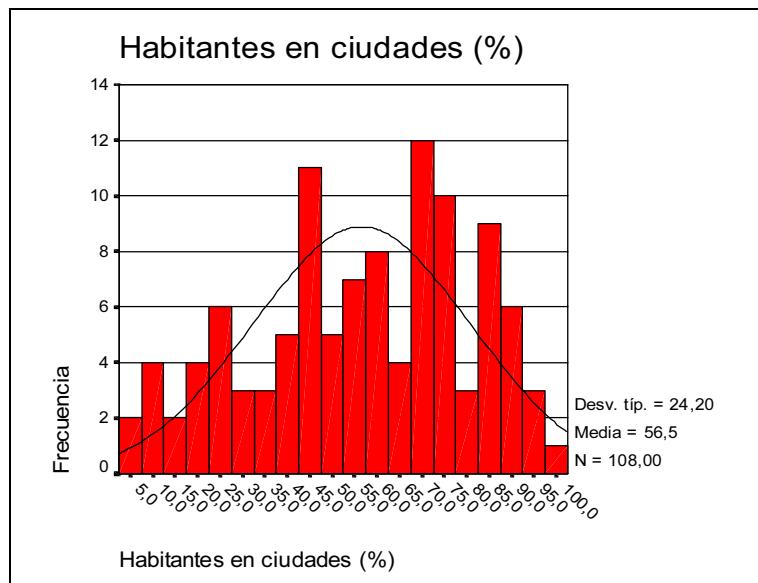


Figura 3-66

EL PROCEDIMIENTO DESCRIPTIVOS DE SPSS

El procedimiento *Descriptivos* muestra estadísticos de resumen univariados para varias variables en una única tabla y calcula valores tipificados (puntuaciones z). Las variables se pueden ordenar por el tamaño de sus medias (en orden ascendente o descendente), alfabéticamente o por el orden en el que se seleccionen las variables (el valor por defecto). En cuanto a estadísticos permite hallar tamaño de muestra, media, mínimo, máximo, desviación típica, varianza, rango, suma, error típico de la media, curtosis y asimetría con sus errores típicos. Cuando se guardan las puntuaciones z, éstas se añaden a los datos del Editor de datos, quedando disponibles para los gráficos, el listado de los datos y los análisis. Cuando las variables se registran en unidades diferentes (por ejemplo, producto interno bruto per cápita y porcentaje de alfabetización), una transformación de puntuación z pondrá las variables en una escala común para una comparación visual más fácil. En cuanto a los datos, utilice variables numéricas después de haberlas inspeccionado gráficamente para registrar errores, valores atípicos y anomalías de distribución. El procedimiento *Descriptivos* es muy eficaz para archivos grandes (de miles de casos).

Para obtener estadísticos descriptivos, elija en los menús *Analizar → Estadísticos descriptivos → Descriptivos* (Figura 3-67) y seleccione una o más variables en la Figura 3-68. Si lo desea, tiene la posibilidad de seleccionar *Guardar valores tipificados como variables* para guardar las puntuaciones z como nuevas variables o pulsar en *Opciones* para seleccionar estadísticos opcionales y el orden de visualización (Figura 3-69). Analizaremos la densidad de población (*densidad*) y el índice de alfabetización total (*alfabet*), masculino (*alfabetm*) y femenino (*alfabetf*) simultáneamente del fichero MUNDO.



Figura 3-67

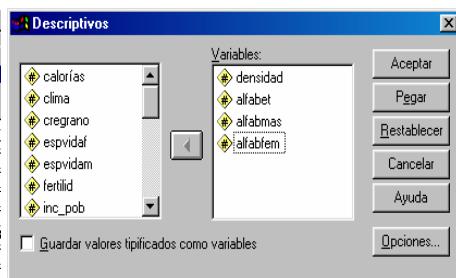


Figura 3-68

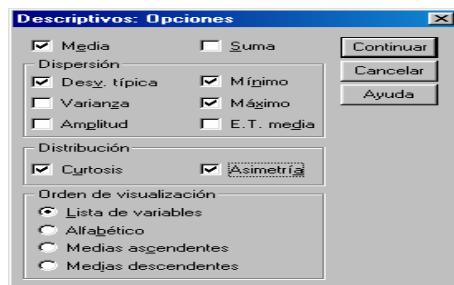


Figura 3-69

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. El botón *Continuar* permite aceptar las asignaciones, el botón *Cancelar* permite ignorarlas y el botón *Pegar* envía la sintaxis del procedimiento a la ventana de sintaxis. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 3-68 para obtener los resultados del procedimiento según se muestra en la Figura 3-70. Se observa que en la parte superior del Visor aparece la sintaxis del procedimiento (dado que se ha activado la opción *Mostrar los comandos en anotaciones* en la solapa *Visor* del cuadro de diálogo *Opciones* del menú *Edición*).

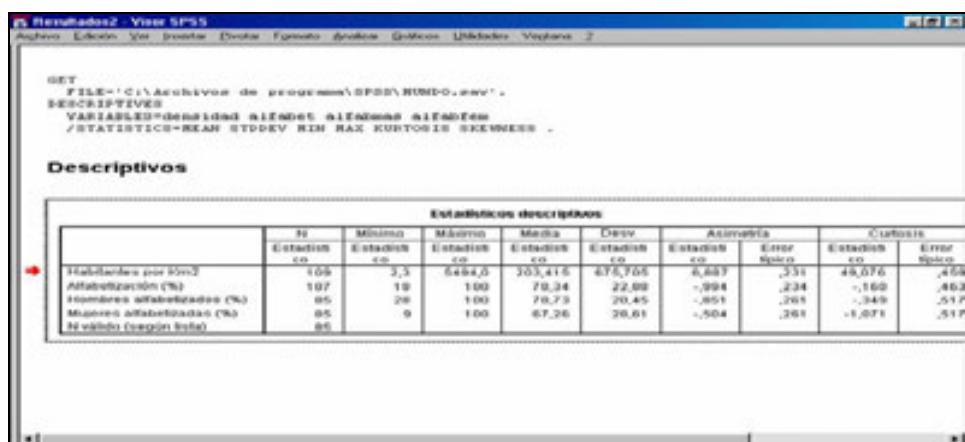


Figura 3-70

LOS PROCEDIMIENTOS INFORME DE ESTADÍSTICOS EN FILAS Y COLUMNAS DE SPSS

El procedimiento Informe de estadísticos en filas genera informes en los cuales se presentan distintos estadísticos de resumen en filas. También se encuentran disponibles listados de los casos, con o sin estadísticos de resumen. Por ejemplo: Una empresa con una cadena de tiendas registra los datos de sus empleados, incluyendo el salario, el cargo, la tienda y la sección en la que trabaja cada uno. Se podría generar un informe que proporcione los datos individuales de cada empleado (listado) desglosados por tienda y sección (variables de ruptura), con estadísticos de resumen (por ejemplo, el salario medio) por tienda, sección y sección dentro de cada tienda.

Para obtener un informe de resumen de estadísticos en filas, elija en los menús *Analizar → Informes → Informe de estadísticos en filas* (Figura 3-71). En *Columnas de datos* seleccione una o más variables para las columnas de datos (Figura 3-72). En el informe se genera una columna para cada variable seleccionada. Para los informes ordenados y mostrados por subgrupos, seleccione una o más variables para *Romper columnas por*. Para los informes con estadísticos de resumen para los subgrupos definidos por las variables de ruptura, seleccione la variable de ruptura de la lista *Romper columnas por* y pulse en *Resumen* en la sección *Romper columnas por*, para especificar las medidas de resumen. Para los informes con estadísticos de resumen globales, pulse en *Resumen*, en la sección *Informe*, para especificar las medidas de resumen.

Romper columnas por muestra una lista de las variables de ruptura opcionales que dividen el informe en grupos y controla los estadísticos de resumen y los formatos de presentación de las columnas de ruptura. Para diversas variables de ruptura, habrá un grupo distinto para cada categoría de cada variable dentro de las categorías de la variable anterior de la lista. Las variables de ruptura deberían ser variables categóricas discretas que dividan los casos en un número limitado de categorías con sentido. Los valores individuales de cada variable de ruptura aparecen ordenados en una columna distinta situada a la izquierda de todas las columnas de datos.

El campo *Informe* de la Figura 3-72 controla las características globales del informe, incluyendo los estadísticos de resumen globales (Figura 3-73), la presentación de los valores perdidos, la numeración de las páginas y los títulos (Figura 3-74). *Mostrar casos* presenta los valores reales (o etiquetas de valor) de las variables de la columna de datos para cada caso, lo que genera un informe a modo de listado, que puede ser mucho más largo que un informe de resumen. *Presentación preliminar* muestra sólo la primera página del informe. Esta opción es útil para hacer un examen preliminar del formato del informe sin tener que procesar el informe completo. Para los informes con variables de ruptura, el archivo de datos se debe ordenar por los valores de estas variables antes de generar el informe.

Si el archivo de datos ya está ordenado por estos valores, se puede ahorrar tiempo de procesamiento seleccionando la opción *Los datos ya están ordenados*. Esta opción es especialmente útil después de generar una presentación preliminar de un informe. El botón *Resumen* permite especificar estadísticos de resumen para las variables de las columnas de datos, dentro de las categorías de las variables de ruptura. El botón *Opciones* (Figura 3-75) permite cambiar el espaciado de las líneas entre las categorías de ruptura o entre los encabezados de las rupturas y los estadísticos de resumen, o mostrar cada categoría de ruptura en una página diferente. El botón *Formato* (Figura 3-76) permite cambiar el espaciado de las líneas entre las categorías de ruptura o entre los encabezados de las rupturas y los estadísticos de resumen, o mostrar cada categoría de ruptura en una página diferente. En la Figura 3-77 se presenta sólo el comienzo de la salida del procedimiento (*Aceptar* en la Figura 3-72).

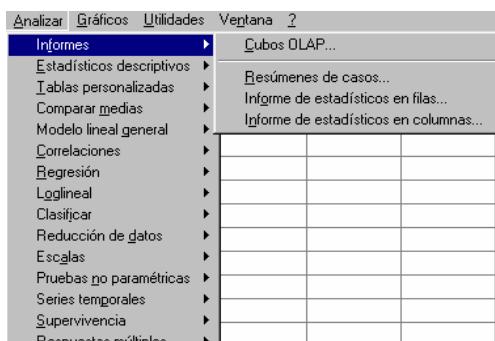


Figura 3-71

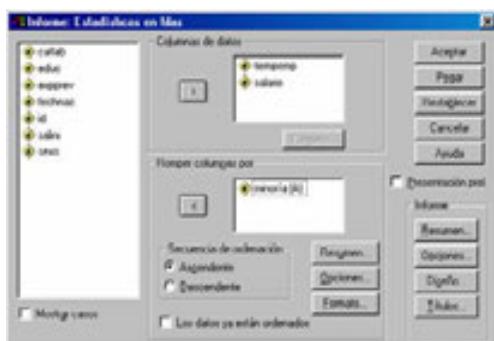


Figura 3-72

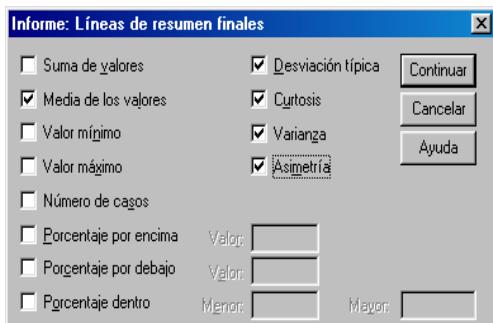


Figura 3-73



Figura 3-74

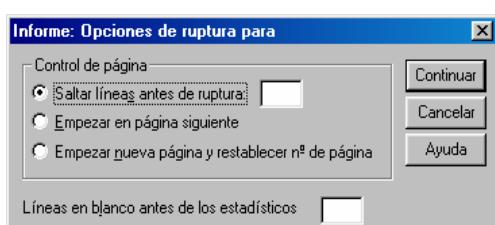


Figura 3-75

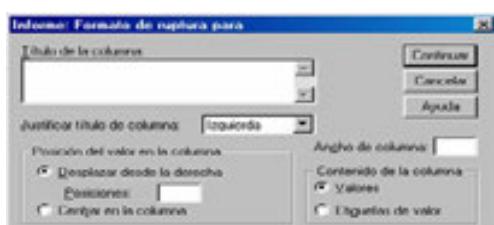


Figura 3-76

Clasificación étnica	Meses desde el contrato	Salario actual
Sí	98	\$57,000
	98	\$40,200
	98	\$21,450
	98	\$21,900
	98	\$45,000
	98	\$32,100
	98	\$36,000
	98	\$21,900
	98	\$27,900
	98	\$24,000

Figura 3-77

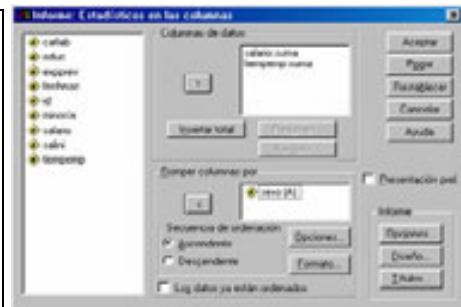


Figura 3-78

El procedimiento *Informe de estadísticos en columnas* (Figura 3-78) es similar al anterior, pero genera informes de resumen en los que diversos estadísticos de resumen aparecen en columnas distintas. *Líneas de resumen* controla el estadístico de resumen mostrado para la variable de las columnas de datos seleccionada. Los estadísticos de resumen disponibles son: suma, media, mínimo, máximo, número de casos, porcentaje de casos por encima o por debajo de un valor especificado, porcentaje de casos dentro de un rango especificado de valores, desviación típica, varianza, curtosis y asimetría. La Figura 3-79 presenta la salida del procedimiento.

Informe		
Página 1		
	Meses	
	Salario actual	desde el contrato
Sexo	Suma	Suma
Hombre	\$10691980	21084
Mujer	\$5,622,895	17362

Figura 3-79

EL PROCEDIMIENTO RESUMIR DE SPSS

El procedimiento *Resumir* calcula estadísticos de subgrupo para las variables dentro de las categorías de una o más variables de agrupación. Se cruzan todos los niveles de las variables de agrupación. Puede elegir el orden en el que se mostrarán los estadísticos. También se muestran estadísticos de resumen para cada variable a través de todas las categorías. Los valores de los datos en cada categoría pueden mostrarse en una lista o suprimirse. Con grandes conjuntos de datos, tiene la opción de listar sólo los primeros *n* casos.

En cuanto a estadísticos, se obtiene la suma, número de casos, media, mediana, mediana agrupada, error típico de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación típica, varianza, curtosis, error típico de curtosis, asimetría, error típico de asimetría, porcentaje de la suma total, porcentaje del N total, porcentaje de la suma en, porcentaje de N en, media geométrica y media armónica.

En cuanto a los datos, las variables de agrupación son variables categóricas cuyos valores pueden ser numéricos o de cadena corta. El número de categorías debe ser razonablemente pequeño. Las otras variables deben poder ordenarse mediante rangos. Algunos de los estadísticos opcionales de subgrupo, como la media y la desviación típica, se basan en la teoría normal y son adecuados para variables cuantitativas con distribuciones simétricas. Los estadísticos robustos, tales como la mediana y el rango, son adecuados para las variables cuantitativas que pueden o no cumplir el supuesto de normalidad.

Para obtener resúmenes de casos, elija en los menús *Anализar* → *Informes* → *Resúmenes de casos* y seleccione las variables en la Figura 3-80. Seleccione una o más variables en la Figura 3-81. Como ejemplo, a partir del fichero MUNDO vamos a clasificar la población mundial (*poblac*), el índice de alfabetización (*alfabet*) y la mortalidad infantil (*mortinf*) por religiones (*relig*). Si lo desea, tiene la posibilidad de seleccionar una o más variables de agrupación para dividir los datos en subgrupos, pulsar en *Opciones* (Figura 3-82) para cambiar el título de los resultados o añadir un texto al pie debajo de los resultados o excluir los casos con valores perdidos, pulsar en *Estadísticos* (Figura 3-83) para acceder a estadísticos adicionales, seleccionar *Mostrar casos* para listar los casos en cada subgrupo. Por defecto, el sistema enumera sólo los 100 primeros casos del archivo. Puede aumentar o disminuir el valor de *LIMITAR los casos a los primeros*, o desactivar ese elemento para enumerar todos los casos. Al pulsar *Aceptar* en la Figura 3-81 se obtiene la salida del procedimiento (Figuras 3-84 y 3-85).



Figura 3-80



Figura 3-81

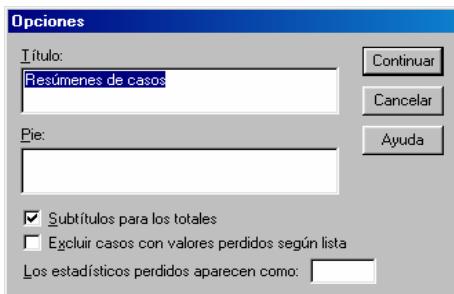


Figura 3-82

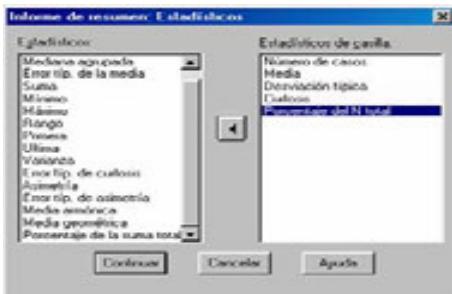


Figura 3-83

Resumen del procesamiento de los casos*						
	Casos					
	Incluidos		Excluidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Alfabetización (%) *						
Religión mayoritaria	97	97,0%	3	3,0%	100	100,0%
Mortalidad infantil (muertes por 1000 nacimientos vivos) *	99	99,0%	1	1,0%	100	100,0%
Religión mayoritaria						
Población x1000 *	99	99,0%	1	1,0%	100	100,0%
Religión mayoritaria						

a. Limitado a los primeros 100 casos.

Figura 3-84

Resúmenes de casos*			
			Mortalidad infantil (muertes por 1000 nacimientos vivos)
			Población x1000
Religión	Animista	N	
mayoritaria	1		118,0
	2	54	77,0
	3	40	113,0
Total		N	2900
		3	3
		Media	102,667
		Desv. tip.	8666,67
		Curtosis	5220,09
		% del total de N	
Budista	1		112,0
	2	99	27,7
	3	77	5,8
	4	99	4,4
	5	93	37,0
	6	91	5,1
Total		N	10000
		6	6
		Media	82,33
		Desv. tip.	40790,67
		Curtosis	45609,74
		% del total de N	
Católica	1		25,6
	2	99	6,7
	3	99	7,2
			8000
			10000

Figura 3-85

Ejercicio 3-1. El fichero HÁBITOS.SAV contiene los resultados simulados de una encuesta realizada a 175 estudiantes relativos a 28 variables de interés. Realizar un análisis exploratorio de datos con las variables CONCIERT que recoge la asistencia anual a conciertos (valores numéricos entre 0 y 26), LECT que recoge los libros leídos anualmente (valores numéricos entre 2 y 24), SEX que recoge el sexo de los encuestados (H=Hombre y M=Mujer) y FÍSICO, que recoge la importancia que se le da al físico (valores Muy poca, Poca, Media, Mucha, y Muchísima, etiquetados de 1 a 5). Posteriormente analizar los libros leídos y la asistencia a los conciertos según la importancia que se la da al físico y por sexos.

Para explorar los datos se comienza estudiando las dos variables cuantitativas. Para ello, elija en los menús *Analizar* → *Estadísticos descriptivos* → *Explorar* y seleccione una o más variables dependientes (Figura 3-86). Si lo desea, tiene la posibilidad de seleccionar una o más variables de factor cuyos valores definirán grupos de casos, (se utilizará más adelante para hacer estudios por sexo e importancia dada al físico), seleccionar una variable de identificación para etiquetar los casos, pulsar en *Estadísticos* para obtener estimadores robustos y valores atípicos (Figura 3-87), pulsar en *Gráficos* para obtener histogramas, pruebas y gráficos de probabilidad normal y diagramas de dispersión por nivel con estadísticos de Levene (Figura 3-88) o pulsar en *Opciones* para manipular los valores ausentes (Figura 3-89). Pulsando *Continuar* en cada Figura, se aceptan sus especificaciones y al pulsar *Aceptar* en la Figura 3-86, se obtiene la salida.

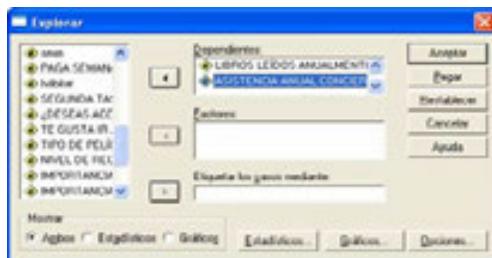


Figura 3-86

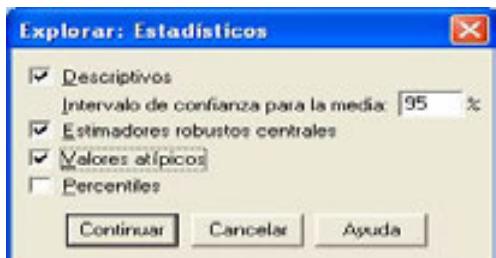


Figura 3-87

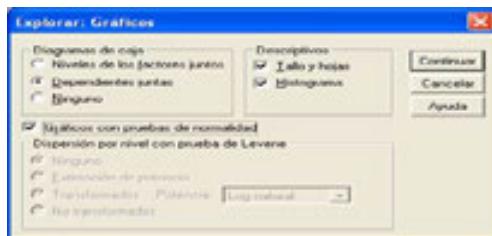


Figura 3-88

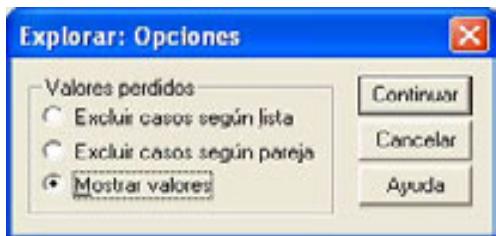


Figura 3-89

En primer lugar se obtiene el resumen de los casos (no hay valores perdidos), estadísticos descriptivos habituales, los M-estimadores muy similares) y valores extremos (no muy exagerados), que no hacen prever la existencia de valores atípicos.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		N	Porcentaje
	N	Porcentaje	N	Porcentaje		
LIBROS LEÍDOS ANUALMENTE	175	100,0%	0	,0%	175	100,0%
ASISTENCIA ANUAL CONCIERTOS	175	100,0%	0	,0%	175	100,0%

Descriptivos

		Estadístico	Error típ.
LIBROS LEÍDOS ANUALMENTE	Media	14,03	,466
	Intervalo de confianza para la media al 95%	Límite inferior	13,12
		Límite superior	14,95
	Media recortada al 5%	14,14	
	Mediana	15,00	
	Varianza	37,941	
	Desv. típ.	6,160	
	Mínimo	2	
	Máximo	24	
	Rango	22	
	Amplitud intercuartil	10,00	
	Asimetría	-,234	,184
	Curtosis	-1,042	,365
	Media	8,55	,556
ASISTENCIA ANUAL CONCIERTOS	Intervalo de confianza para la media al 95%	Límite inferior	7,45
		Límite superior	9,65
	Media recortada al 5%	8,11	
	Mediana	6,00	
	Varianza	54,077	
	Desv. típ.	7,354	
	Mínimo	0	
	Máximo	26	
	Rango	26	
	Amplitud intercuartil	10,00	
	Asimetría	,961	,184
	Curtosis	-,081	,365

Estimadores-M

	Estimador-M de Huber(a)	Biponderado de Tukey(b)	Estimador-M de Hampel(c)	Onda de Andrews(d)
LIBROS LEÍDOS ANUALMENTE	14,38	14,31	14,19	14,31
ASISTENCIA ANUAL CONCIERTOS	7,00	5,96	7,04	5,95

a La constante de ponderación es 1,339.

b La constante de ponderación es 4,685.

c Las constantes de ponderación son 1,700, 3,400 y 8,500.

d La constante de ponderación es 1,340*pi.

Valores extremos

			Número del caso	Valor
LIBROS LEÍDOS ANUALMENTE	Mayores	1	27	24
		2	40	24
		3	67	24
		4	80	24
		5	107	24(a)
	Menores	1	153	2
		2	142	2
		3	113	2
		4	102	2
		5	73	2(b)
ASISTENCIA ANUAL CONCIERTOS	Mayores	1	19	26
		2	59	26
		3	99	26
		4	139	26
		5	6	24(a)
	Menores	1	167	0
		2	161	0
		3	154	0
		4	141	0
		5	127	0(c)

a En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 24.

b En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 2.

c En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 0.

Los p-valores de los contrastes de normalidad, así como los diagramas de tallos y hojas, que se presentan a continuación, muestran la normalidad de los datos.

Pruebas de normalidad

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
LIBROS LEÍDOS ANUALMENTE	,117	175	,000	,953	175	,000
ASISTENCIA ANUAL CONCIERTOS	,190	175	,000	,876	175	,000

a Corrección de la significación de Lilliefors

LIBROS LEÍDOS ANUALMENTE Stem-and-Leaf Plot

Frequency Stem & Leaf

,00	0 .
8,00	0 . 22222222
12,00	0 . 555555555555
12,00	0 . 666666667777
16,00	0 . 8888999999999999
16,00	1 . 000000000001111
16,00	1 . 22222223333333
10,00	1 . 444445555
19,00	1 . 6666666667777777
27,00	1 . 88888888888888899999999
23,00	2 . 0000011111111111111111
8,00	2 . 2222333
8,00	2 . 4444444

ASISTENCIA ANUAL CONCIERTOS Stem-and-Leaf Plot

Frequency Stem & Leaf

Los histogramas de las Figuras 3-90 y 3-91, así como los gráficos Q-Q de las Figuras 3-92 y 3-93 corroboran la normalidad.

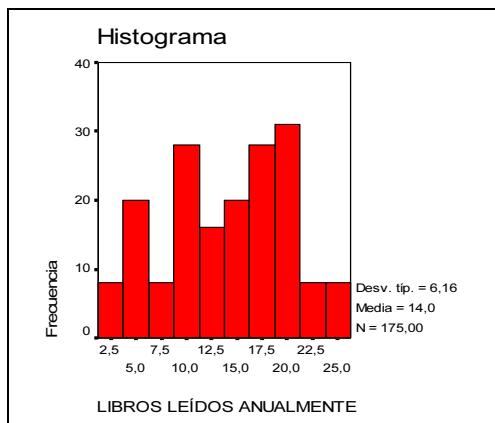


Figura 3-90

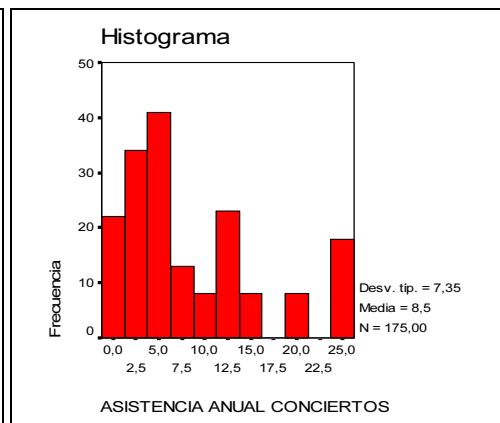


Figura 3-91

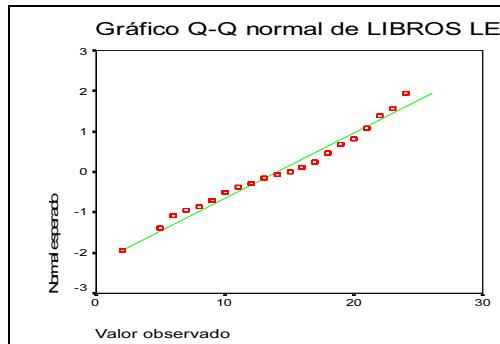


Figura 3-92

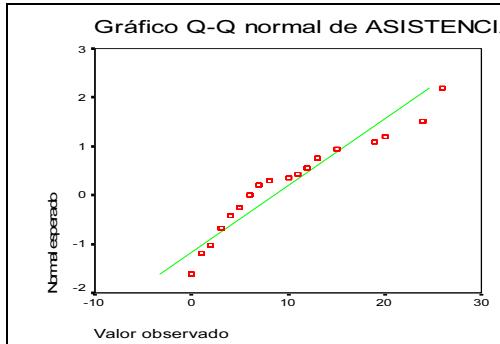


Figura 3-93

El gráfico de caja y bigotes (Figura 3-94) muestra la inexistencia de valores atípicos, una ligera asimetría hacia la izquierda de la variable CONCIERT y hacia la derecha de la variable FÍSICO, observadas ya en los histogramas y gráficos de tallo y hojas.

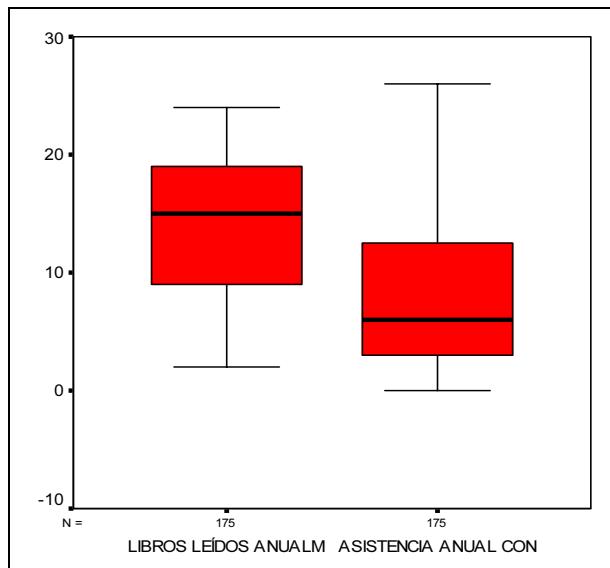


Figura 3-94

Para realizar un análisis exploratorio de los libros leídos anualmente y la asistencia anual a los conciertos clasificados por sexo e importancia del físico, se rellena la pantalla del procedimiento explorar como se indica en la Figura 3-95. Las pantallas *Gráficos* y *Estadísticos* se llenan según se muestra en las Figuras 3-96 y 3-97. Al pulsar en *Aceptar* se obtiene el análisis exploratorio según categorías de las variables cualitativas.



Figura 3-95

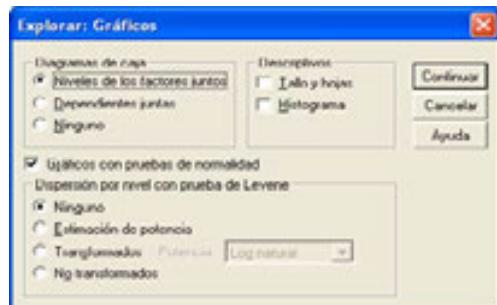


Figura 3-96

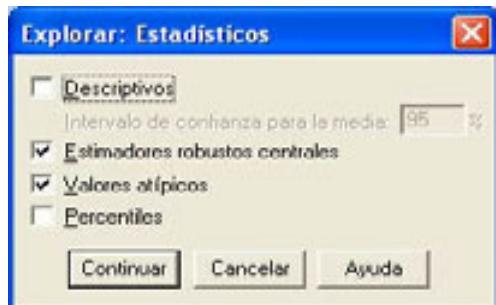


Figura 3-97

A continuación se presentan los datos por categorías de sexo. No hay datos perdidos ni atípicos claros y se observa que se cumple la normalidad por categorías de sexo.

Resumen del procesamiento de los casos

	SEXO	Casos					
		Válidos		Perdidos		Total	
		N	Porcentaje	N	Porcentaje	N	Porcentaje
LIBROS LEÍDOS ANUALMENTE	HOMBRE	77	100,0%	0	,0%	77	100,0%
ASISTENCIA ANUAL CONCIERTOS	MUJER	98	100,0%	0	,0%	98	100,0%
	HOMBRE	77	100,0%	0	,0%	77	100,0%
	MUJER	98	100,0%	0	,0%	98	100,0%

Estimadores-M

	SEXO	Estimador-M de Huber(a)	Biponderado de Tukey(b)	Estimador-M de Hampel(c)	Onda de Andrews(d)
LIBROS LEÍDOS ANUALMENTE	HOMBRE	12,33	12,35	12,37	12,35
ASISTENCIA ANUAL CONCIERTOS	MUJER	16,16	16,37	15,88	16,38
	HOMBRE	4,76	4,10	4,54	4,08
	MUJER	9,45	9,64	9,90	9,65

a La constante de ponderación es 1,339.

b La constante de ponderación es 4,685.

c Las constantes de ponderación son 1,700, 3,400 y 8,500.

d La constante de ponderación es 1,340*pi.

Pruebas de normalidad

	SEXO	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
LIBROS LEÍDOS ANUALMENTE	HOMBRE	,105	77	,036	,961	77	,018
	MUJER	,149	98	,000	,930	98	,000
ASISTENCIA ANUAL CONCIERTOS	HOMBRE	,251	77	,000	,860	77	,000
	MUJER	,153	98	,000	,897	98	

a Corrección de la significación de Lilliefors

A continuación se presentan los datos por categorías de importancia dada al físico. No hay datos ausentes, los M-estimadores no presagian de modo claro datos atípicos y se observa que se cumple la normalidad por categorías (salvo en un par de casos).

Resumen del procesamiento de los casos

	importancia física	Casos					
		Válidos		Perdidos		Total	
		N	Porcentaje	N	Porcentaje	N	Porcentaje
LIBROS LEÍDOS ANUALMENTE	MUY POCA	17	100,0%	0	,0%	17	100,0%
	POCA	17	100,0%	0	,0%	17	100,0%
	MEDIA	43	100,0%	0	,0%	43	100,0%
	MUCHA	52	100,0%	0	,0%	52	100,0%
	MUCHÍSIMA	46	100,0%	0	,0%	46	100,0%
ASISTENCIA ANUAL CONCIERTOS	MUY POCA	17	100,0%	0	,0%	17	100,0%
	POCA	17	100,0%	0	,0%	17	100,0%
	MEDIA	43	100,0%	0	,0%	43	100,0%
	MUCHA	52	100,0%	0	,0%	52	100,0%
	MUCHÍSIMA	46	100,0%	0	,0%	46	100,0%

Estimadores-M

	importancia por físico	Estimador-M de Huber(a)	Biponderado de Tukey(b)	Estimador-M de Hampel(c)	Onda de Andrews (d)
LIBROS LEÍDOS ANUALMENTE	MUY POCA	12,75	12,86	12,82	12,86
	POCA	14,70	14,69	14,55	14,69
	MEDIA	12,83	12,88	12,96	12,88
	MUCHA	13,78	13,70	13,50	13,70
	MUCHÍSIMA	17,06	17,28	16,70	17,29
ASISTENCIA ANUAL CONCIERTOS	MUY POCA	5,60	5,53	5,70	5,53
	POCA	4,33	3,92	3,97	3,92
	MEDIA	7,45	7,52	7,87	7,53
	MUCHA	7,04	6,56	7,08	6,55
	MUCHÍSIMA	9,46	9,37	10,03	9,39

Pruebas de normalidad

	importancia por físico	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
LIBROS LEÍDOS ANUALMENTE	MUY POCA	,128	17	,200(*)	,906	17	,086
	POCA	,121	17	,200(*)	,967	17	,764
	MEDIA	,134	43	,050	,957	43	,104
	MUCHA	,141	52	,011	,937	52	,008
	MUCHÍSIMA	,159	46	,005	,893	46	,001
ASISTENCIA ANUAL CONCIERTOS	MUY POCA	,152	17	,200(*)	,921	17	,156
	POCA	,361	17	,000	,697	17	,000
	MEDIA	,231	43	,000	,893	43	,001
	MUCHA	,196	52	,000	,889	52	,000
	MUCHÍSIMA	,201	46	,000	,853	46	,000

Los gráficos de caja y bigotes de las Figuras 3-98 y 3-99 muestran que sólo se presentan casos atípicos para la asistencia anual a los conciertos de los estudiantes que le dan poca importancia al físico

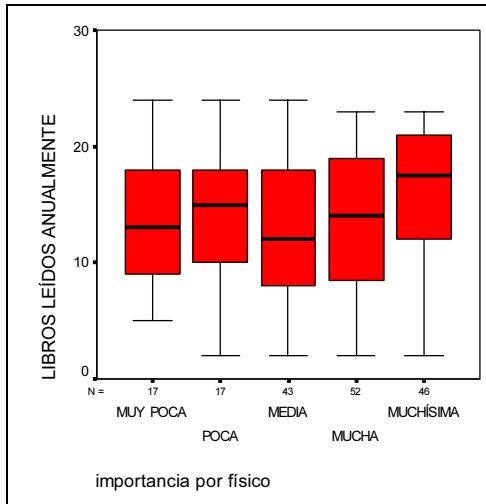


Figura 3-98

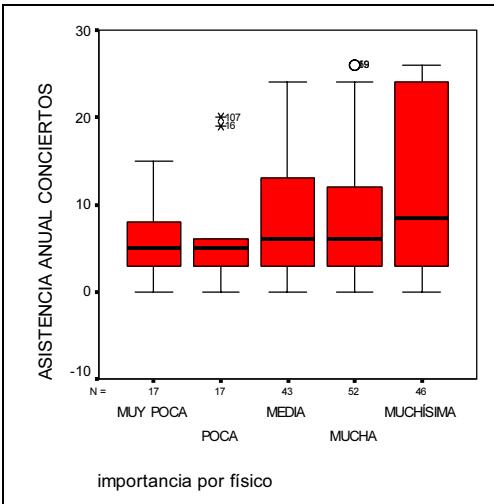


Figura 3-99

Los gráficos de caja y bigotes de las Figuras 3-100 y 3-101 muestran que sólo se presentan casos atípicos para la asistencia anual a los conciertos de los estudiantes hombres.

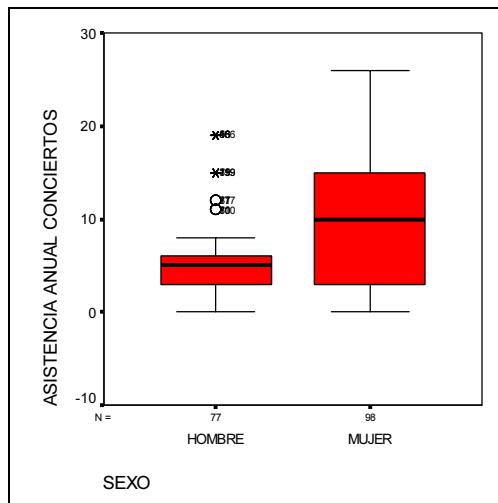


Figura 3-100

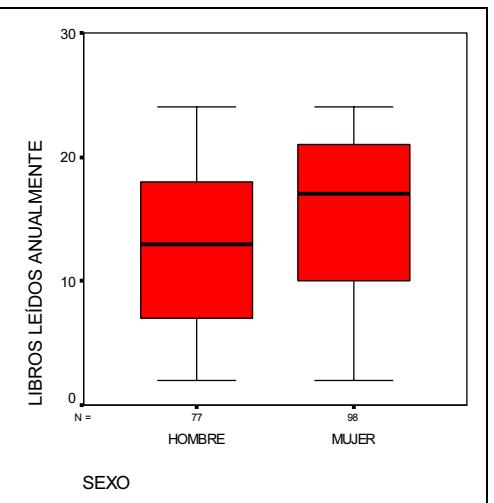


Figura 3-101

Ejercicio 3-2. El fichero **AUSENTES.xls** contiene los resultados simulados de una encuesta realizada a 70 empresas en la que se midieron 14 variables. Las 10 primeras variables son cuantitativas y recogen información sobre velocidad de entrega del producto, nivel de precios, flexibilidad de precios, imagen del fabricante, servicio conjunto, imagen de fuerza de ventas, calidad del producto, nivel de fidelidad y nivel de satisfacción respectivamente. Se trata de realizar el análisis y tratamiento de los datos ausentes y de los datos atípicos para las 10 variables cuantitativas previo a cualquier otro análisis formal a llevar a cabo. Aunque la variable **V8** puede ser cualitativa vamos a considerarla cuantitativa.

Como la información viene dada en un fichero Excel, comenzaremos transformándola a fichero SPSS. Para ello elegimos en los menús: *Archivo → Abrir → Datos* (Figura 3-102). En el cuadro de diálogo *Abrir archivo*, seleccione el tipo de archivo y el archivo que desea abrir (Figura 3-103) y pulse en *Abrir*. Si la primera fila de la hoja de cálculo contiene encabezados o etiquetas de columna, pulse en *Leer nombres de variable en la primera fila de datos* (Figura 3-104). Si los datos que desea leer no comienzan en la primera fila de la hoja de cálculo, introduzca el *Rango* de casillas en el cuadro de diálogo *Apertura de fuente de datos*. En los archivos de Excel 5 o de versiones posteriores, si los datos que desea leer no se encuentran en la primera hoja del archivo, seleccione la hoja que desea leer en el campo *Hoja de trabajo*. Si los datos aparecen como valores perdidos o se leen como datos de cadena, probablemente habrá intentado leer la primera fila como si se tratara de datos, cuando en realidad contiene encabezados. Se obtienen ya los datos en fichero SPSS (Figura 3-105).

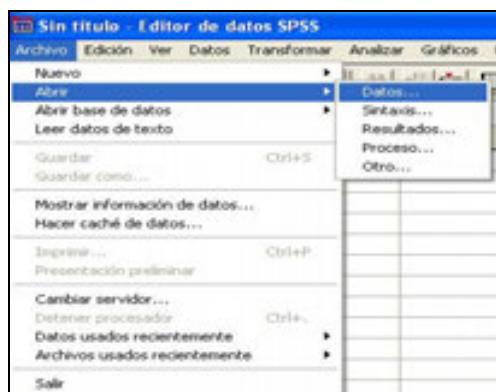


Figura 3-102

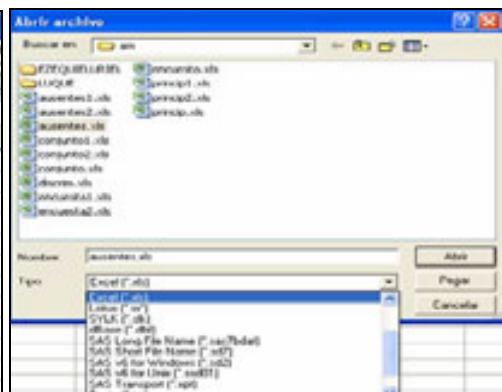


Figura 3-103

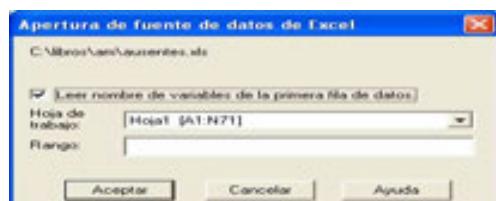


Figura 3-104

	v1	v2	v3	v4	
1	3,3	.	8,6	.	4
2	.	.	4	.	3
3	3,0	.	.	9,1	7
4	.	.	1,5	.	5
5	5,1	1,4	.	.	5
6	4,6	2,1	7,9	.	6
7	.	1,5	.	.	5
8	5,2	1,3	9,7	.	6

Figura 3-105

Tras observar la presencia de datos ausentes en una distribución será necesario detectar si éstos se distribuyen aleatoriamente. Habrá que detectar que el efecto de los datos ausentes es importante mediante pruebas formales de aleatoriedad.

Ya sabemos que una *primera prueba para valorar los datos ausentes* es la *prueba de las correlaciones dicotomizadas*. Para realizar esta prueba, para cada variable V del análisis se construye una variable dicotomizada asignando el valor cero a los valores ausentes y el valor uno a los valores presentes. A continuación se halla la matriz de correlaciones de las variables cuantitativas dicotomizadas acompañada de los contrastes de significatividad de cada coeficiente de correlación de la matriz. Si los elementos de la matriz de correlaciones son no significativos, los datos ausentes son completamente aleatorios. Si existe alguna correlación significativa y la mayor parte son no significativas, los datos ausentes pueden considerarse aleatorios. En ambos casos podrán realizarse análisis estadísticos previa imputación de la información faltante.

Comenzamos generando las variables D1 a D10 de modo que D_i vale 0 para valores ausentes de V_i , y vale 1 para valores presentes. Para ello se utiliza *Datos* → *Seleccionar casos* (Figura 3-106) y en la pantalla *Seleccionar casos* se elige *Si se satisface la condición* y se hace clic en *Sí* (Figura 3-107). En la Figura 3-108 se introduce la función $1-\text{MISSING}(V1)$ que calculará D1. Al hacer clic en *Continuar*, la nueva variable ya aparece en la pantalla *Seleccionar casos* (Figura 3-109). Al hacer clic en *Aceptar*, la nueva variable ya se incorpora al fichero SPSS con nombre *filter\$* (Figura 3-110). Se pasa a vista variables y se asigna el nuevo nombre D1 (Figura 3-111). Al volver a vista de datos ya se observa la nueva variable (Figura 3-112). Repitiendo el mismo proceso, se generan las variables D2 a D10 (Figura 3-113).



Figura 3-106

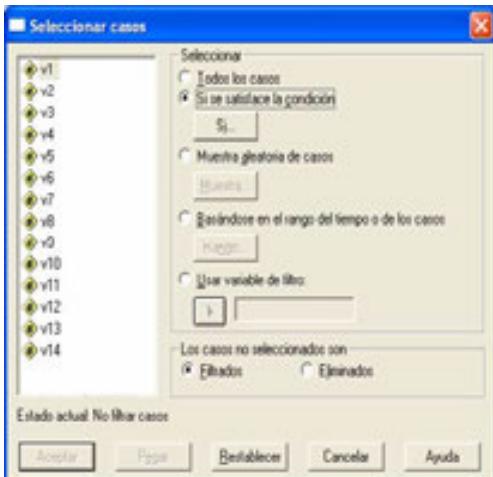


Figura 3-107

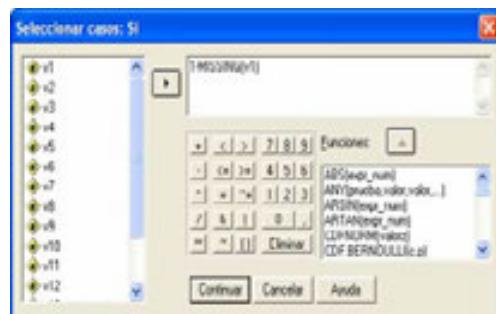


Figura 3-108

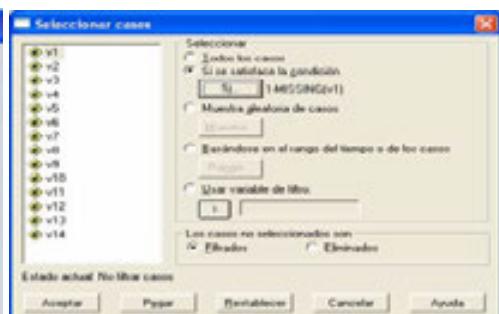


Figura 3-109

AUSENTES.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Análisis Gráficos Utilidades Ayuda

Vista de datos Vista de variables

SPSS El procesador está preparado Filtro

v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	Filas	v1	v2
1	2.1	1.8	0.3	0	41	4.5	1	0	0	0	0	0	2	1	0	0
2	1.2	1.7	6.2	0	36	3.3	1	0	0	0	0	0	2	0	0	1
3	3.6	3.4	0	0	55	5.2	1	1	0	0	0	0	2	1	0	0
4	1.9	2.6	7.2	1	36	0	0	0	0	0	0	0	2	0	0	0
5	2.3	2.6	3.0	0	49	4.9	1	0	0	0	0	0	2	1	0	0
6	3.4	2.8	4.7	0	39	5.9	1	0	0	0	0	0	2	0	0	1
7	1.9	2.5	7.2	1	36	0	0	0	0	0	0	0	2	0	0	0
8	3.2	3.9	6.7	0	54	5.8	1	0	0	0	0	0	2	1	0	0
9	3.1	1.7	5.6	0	49	5.4	1	0	0	0	0	0	2	0	0	1
10										0	0	0	0	0	0	0
11	3.0	3.8	7.0	0	49	4.8	1	0	0	0	0	0	2	1	0	0
12	3.3	1.4	6.8	0	39	6.5	1	0	0	0	0	0	2	0	0	1
13	3.6	4.0	5.9	0	49	5.2	1	0	0	0	0	0	2	1	0	0
14	2.2								0	0	0	0	0	0	0	0
15	3.2	2.1	6.0	0	31	4.0	1	0	0	0	0	0	2	1	0	0
16		2.1	8.4	1	26	3.4	0	0	0	0	0	0	2	0	0	0
17	3.3	2.8	7.1	0	60	5.2	1	0	0	0	0	0	2	1	0	0
18		2.7	6.4	1	36	3.7	0	0	0	0	0	0	2	0	0	0
19	2.6	2.9		1		4.3	0	0	0	0	0	0	2	0	0	0
20	3.2	3.7	8.0	0	55	6.4	1	0	0	0	0	0	2	1	0	0
21	2.0	2.8		0	32	4.3	1	0	0	0	0	0	2	0	0	0
22		2.8	9.3	0	47	9.0	1	0	0	0	0	0	2	1	0	0
23	2.1	1.4	6.6	1	39	4.4	0	0	0	0	0	0	2	0	0	0
24		2.7	3.7	0	65	6.0	1	0	0	0	0	0	2	0	0	0
25	3.0	2.6	6.8	0	54	5.9	1	0	0	0	0	0	2	1	0	0
26		2.7	1.7	4.8	0	49	4.7	1	0	0	0	0	2	1	0	0
27	3.6	2.9	6.2	0		4.4	1	0	0	0	0	0	2	1	0	0
28		3.4	1.5	6.9	0	40	6.6	1	0	0	0	0	2	0	0	0
29	3.9	9.8	0	0	54	5.9	1	0	0	0	0	0	2	1	0	0
30	3.0	2.6	8.0	0	55	6.0	1	0	0	0	0	0	2	1	0	0

Figura 3-110

AUSENTES.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Análisis Gráficos Utilidades Ayuda

Vista de datos Vista de variables

SPSS El procesador está preparado Filtro

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	v1	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
2	v2	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
3	v3	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
4	v4	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
5	v5	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
6	v6	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
7	v7	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
8	v8	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
9	v9	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
10	v10	Numerico	11	1		Ninguno	Ninguno	8	Derecha	Escala
11	v11	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
12	v12	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
13	v13	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
14	v14	Numerico	11	0		Ninguno	Ninguno	8	Derecha	Escala
15	v15	Numerico	1	0	-1.MENOSUNO(1) / 0, NO_NINGUNO(Ninguno)	Ninguno	Ninguno	8	Derecha	Escala

Figura 3-111

A screenshot of the SPSS Data Editor window titled "AUSENTES.sav - Editor de datos SPSS". The window shows a data table with 16 rows and 14 columns. The columns are labeled v5 through v14 and d1. The data contains various numerical values (e.g., 1.2, 3.5, 2.5, 3.3, 2.8, 3.4, 2.5, 3.2, 3.9, 3.1, 3.0, 3.3, 3.4, 2.5, 2.2, 2.2) and some missing values represented by periods. The last row (row 16) contains mostly missing values.

Figura 3-112

A screenshot of the SPSS Data Editor window titled "AUSENTES.sav - Editor de datos SPSS". The window shows a data table with 16 rows and 14 columns. The columns are labeled v13 through v14 and d1 through d10. The data contains various numerical values (e.g., 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15) and some missing values represented by periods. The last row (row 16) contains mostly missing values.

Figura 3-113

Para obtener la matriz de correlaciones bivariadas elegimos en los menús *Analizar* → *Correlaciones* → *Bivariadas* (Figura 3-114) y seleccionamos las variables numéricas D1 a D10 (Figura 3-115). Al pulsar en *Aceptar*, se obtiene la matriz de correlaciones de la Figura 3-116. En esta matriz los valores significativos al 95% están señalados con un asterisco y los valores significativos al 99% están señalados con dos asteriscos. Se observa que hay 19 valores significativos de entre 10 (casi un 20%), lo que indica que los valores no significativos son muy dominantes en la matriz de correlaciones. La última fila y columna de la matriz, no aparece, debido al ruido introducido por la variable V8 que se consideró cuantitativa con sólo dos valores diferentes. Con ciertas reservas, podría considerarse que los datos ausentes se distribuyen aleatoriamente. Podemos entonces aplicar técnicas de imputación de la información faltante.



Figura 3-114



Figura 3-115

Correlaciones									
1-MISSING(v 1) (FILTER)	1-MISSING(v 2) (FILTER)	1-MISSING(v 3) (FILTER)	1-MISSING(v 4) (FILTER)	1-MISSING(v 5) (FILTER)	1-MISSING(v 6) (FILTER)	1-MISSING(v 7) (FILTER)	1-MISSING(v 8) (FILTER)	1-MISSING(v 9) (FILTER)	
1	-.115 .367 64 -.115 .367 64	.214 .090 64 1 .375 64	.115 .366 64 -.113 .375 64	.101 .427 64 .053 .675 64	.018 .890 64 -.027 .833 64	.065 .612 64 .291* .020 64	-.117 .358 64 .203 .107 64	.065 .612 64 .156 .218 64	.065 .612 64 .329** .008 64
.214 .090 64	-.113 .375 64	1 .329 64	-.124 .329 64	.216 .086 64	.083 .515 64	.182 .151 64	-.086 .498 64	.061 .835 64	.061
.115 .366 64	.053 .675 64	-.124 .329 64	1 .329 64	.116 .360 64	.561** .000 64	.293* .019 64	.325** .009 64	.293*	.293*
.101 .427 64	-.027 .833 64	.216 .086 64	.116 .360 64	1 .209 64	.159 .209 64	.170 .179 64	.225 .074 64	.170 .179 64	.170
.018 .890 64	.291* .020 64	.083 .515 64	.554** .000 64	.159 .209 64	1 .000 64	.587** .000 64	.385** .002 64	.587** .000 64	.587**
.065 .612 64	.203 .107 64	.182 .151 64	.293* .019 64	.170 .179 64	.587** .000 64	1 .107 64	.204 .107 64	.571** .000 64	.571**
-.117 .358 64	.156 .218 64	-.086 .498 64	.325** .009 64	.225 .074 64	.385** .007 64	.204 .107 64	1 .107 64	.204 .107 64	.204
.065 .612 64	.329** .008 64	.061 .635 64	.293* .019 64	.170 .179 64	.587** .000 64	.571** .000 64	.204 .107 64	.204 .107 64	1 .

Figura 3-116

Una segunda prueba para valorar los datos ausentes para una única variable Y consiste en formar dos grupos de valores para Y, los que tienen datos ausentes y los que no los tienen. A continuación, para cada variable X distinta de Y, se realiza un test para determinar si existen diferencias significativas entre los dos grupos de valores determinados por la variable Y (ausentes y no ausentes) sobre X.

Si vamos considerando como Y cada una de las variables del análisis y repitiendo el proceso anterior se encuentra que todas las diferencias son no significativas, se puede concluir que los datos ausentes obedecen a un *proceso completamente aleatorio* y por tanto pueden realizarse análisis estadísticos fiables con nuestras variables *imputando los datos ausentes* por los métodos que se verán más adelante. Si un porcentaje bastante alto de las diferencias son no significativas, puede considerarse que los datos ausentes obedecen a un *proceso aleatorio* (no completamente aleatorio) que también permitirá realizar análisis estadísticos fiables con nuestras variables previa *imputación de la información faltante*, aunque con menos fiabilidad que en el caso anterior.

Comenzamos considerando los dos grupos formados en la variable V1 (valores válidos y valores ausentes) que vienen definidos por la variable D1 y hacemos un contraste de igualdad de medias para los dos grupos de valores definidos en cada una de las restantes variables (V2 a V10) por los valores de D1. Elegimos en los menús *Analizar → Comparar medias → Prueba T para muestras independientes* (Figura 3-117). Seleccionamos las variables de contraste cuantitativas V2 a V10 en el campo *Variables* (Figura 3-118) para calcular una prueba T diferente para cada variable. Seleccionamos la variable D1 como variable de agrupación. Pulsamos en *Definir grupos* para especificar dos códigos para los grupos que desea comparar definidos por los valores de la variable de agrupación (Figura 3-119). Si lo desea, puede pulsar en *Opciones* para controlar el tratamiento de los datos perdidos y el nivel del intervalo de confianza (Figura 3-120). Al pulsar *Aceptar* en la Figura 3-118 se obtienen un resumen de estadísticos para la muestra y los resultados del contraste de la T titulados Prueba de muestras independientes.

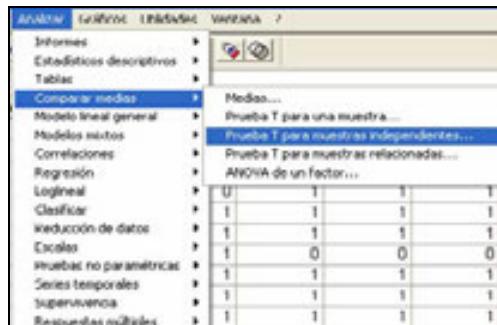


Figura 3-117



Figura 3-118

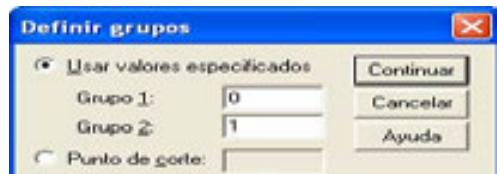


Figura 3-119

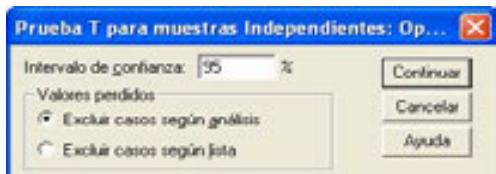


Figura 3-120

Prueba de muestras independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gr	Sig. (bilateral)	Diferencia de medias	Error tip. de la diferencia	95% Intervalo de confianza para la diferencia	
								Inférieur	Superior
V2 Se han asumido varianzas iguales No se han asumido varianzas iguales	,036	,960	,196	61	,865	,062	,2649	-,4797	,6937
V3 Se han asumido varianzas iguales No se han asumido varianzas iguales	,187	,869	-,559	50	,579	-,374	,6699	-,1199	,9711
V4 Se han asumido varianzas iguales No se han asumido varianzas iguales	2,702	,106	-1,270	58	,209	-,53	,420	-,1,376	,308
V5 Se han asumido varianzas iguales No se han asumido varianzas iguales	8,518	,005	-3,151	55	,003	-,858	,2007	-,1,0759	,2393
V6 Se han asumido varianzas iguales No se han asumido varianzas iguales	,000	,779	-1,740	59	,087	-,346	,1989	-,7440	,0520
V7 Se han asumido varianzas iguales No se han asumido varianzas iguales	1,663	,293	1,037	54	,204	-,525	,5061	-,4096	1,5396
V8 Se han asumido varianzas iguales No se han asumido varianzas iguales	2,730	,104	3,006	60	,000	,47	,119	,237	,713
V9 Se han asumido varianzas iguales No se han asumido varianzas iguales	3,808	,056	-2,613	54	,012	-,6,99	2,677	-,12,361	,1,629
V10 Se han asumido varianzas iguales No se han asumido varianzas iguales	,446	,506	-1,894	62	,063	-,517	,2728	-,1,0820	,0296

Se observa que salvo para las variables V5 y V8, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de V1 en las variables V2, V3, V4, V6, V7, V9 y V10 (los intervalos de confianza contienen el valor cero). El contraste de igualdad de medias se realiza suponiendo varianzas iguales (primera línea de la tabla para cada variable) y desiguales (segunda línea para cada variable).

Ahora consideramos los dos grupos formados en la variable V2 (valores válidos y valores ausentes) que vienen definidos por la variable D2 y hacemos un contraste de igualdad de medias para los dos grupos de valores definidos en cada una de las restantes variables (V1 y V3 a V10) por los valores de D2.

A continuación se observa el resultado que muestra que para todas las variables (excepto V4, V5 y V8), no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de V2 en cada una de ellas (los intervalos de confianza contienen el valor cero). Repitiendo los contrastes de igualdad de medias para los grupos que determinan los valores válidos y ausentes de las variables V3 a V10 en el resto de las variables, tenemos los resultados que se presentan en las tablas siguientes (sólo de V3 a V5):

Prueba de contrastes independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						95% Intervalo de confianza para la diferencia		
			r	tug	t	gi	tug (bilateral)	Diferencia en medias	Error tig. de la diferencia	Intervalo anterior	Intervalo superior
V3	Se han assumido varianzas iguales		17,312	,000	-1,929	50	,059	-1,374	,7123	-2,8048	,0563
	No se han assumido varianzas iguales				-1,162	9,853	,273	-1,374	1,1831	-4,0233	1,2747
V4	Se han assumido varianzas iguales		,005	,945	2,351	59	,022	1,16	,493	,172	2,144
	No se han assumido varianzas iguales				2,639	14,574	,016	1,16	,439	,220	2,096
V5	Se han assumido varianzas iguales		2,694	,113	3,761	66	,000	,897	,2391	,4177	1,3769
	No se han assumido varianzas iguales				4,396	16,929	,006	,897	,2041	,4646	1,3296
V6	Se han assumido varianzas iguales		,084	,362	2,575	59	,013	,839	,2407	,1417	1,3370
	No se han assumido varianzas iguales				2,276	16,247	,020	,839	,2699	,0429	1,2291
V7	Se han assumido varianzas iguales		,249	,620	1,095	54	,270	,715	,6526	-,5930	2,0220
	No se han assumido varianzas iguales				1,056	9,230	,218	,715	,6769	,0107	2,2399
V8	Se han assumido varianzas iguales		,311	,579	,321	60	,749	,05	,168	,292	,399
	No se han assumido varianzas iguales				,305	12,185	,765	,05	,176	,330	,437
V9	Se han assumido varianzas iguales		,129	,730	,975	54	,934	,779	,947	-2,963	11,465
	No se han assumido varianzas iguales				,924	7,668	,984	,779	,4059	,5794	13,292
V10	Se han assumido varianzas iguales		,736	,394	-,300	62	,765	,102	,3395	,7806	,5768
	No se han assumido varianzas iguales				-,204	11,240	,842	,102	,493	,1,1980	,8943
V11	Se han assumido varianzas iguales		1,369	,248	,352	43	,802	,892	,3631	,6406	,8239
	No se han assumido varianzas iguales				,213	10,417	,835	,892	,4294	,8600	1,0434

Prueba de contrastes independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						95% Intervalo de confianza para la diferencia		
			r	tug	t	gi	tug (bilateral)	Diferencia en medias	Error tig. de la diferencia	Intervalo anterior	Intervalo superior
V4	Se han assumido varianzas iguales		0,083	,017	-3,775	59	,000	-1,62	,430	-2,684	,301
	No se han assumido varianzas iguales				-2,738	12,084	,017	-1,62	,593	-2,005	,341
V5	Se han assumido varianzas iguales		4,092	,048	-1,846	56	,070	,501	,2712	-1,0445	,0428
	No se han assumido varianzas iguales				-1,378	9,206	,200	,501	,3634	-1,3186	,3172
V6	Se han assumido varianzas iguales		,669	,617	,626	59	,534	,151	,2411	,6334	,3316
	No se han assumido varianzas iguales				,695	16,712	,498	,151	,2170	,6693	,3075
V7	Se han assumido varianzas iguales		2,158	,148	,639	54	,626	,460	,6263	-1,6581	,8662
	No se han assumido varianzas iguales				,515	9,894	,618	,460	,7770	-2,1391	1,3081
V8	Se han assumido varianzas iguales		,657	,421	,491	60	,625	,08	,156	,235	,399
	No se han assumido varianzas iguales				,469	15,879	,645	,08	,163	,270	,423
V9	Se han assumido varianzas iguales		,174	,679	-1,040	54	,303	,345	,3318	-10,102	3,202
	No se han assumido varianzas iguales				-1,173	16,267	,359	,345	,2,942	,9,711	,8,811
V10	Se han assumido varianzas iguales		3,594	,063	-2,396	62	,020	,750	,3143	-1,2782	,1217
	No se han assumido varianzas iguales				-1,698	12,539	,122	,750	,4,622	-1,7397	,2397
V11	Se han assumido varianzas iguales		1,356	,251	,048	43	,962	,021	,4276	,8418	,8828
	No se han assumido varianzas iguales				,080	8,185	,953	,021	,3409	,7628	,8038
V2	Se han assumido varianzas iguales		,164	,687	-1,026	54	,316	,310	,3018	,9157	,2962
	No se han assumido varianzas iguales				-1,041	15,949	,314	,310	,2977	,9409	,3214

Prueba de varianzas independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						95% Intervalo de confianza para la diferencia	
	F	Sig.	t	gr	Sig. (bilateral)	Diferencia de medias	Error std. de la diferencia			
								inferior	superior	
V5 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,403	,528	,415	55	,683	,187	,4557	-1,1002	,3261	
				,508	,2371	,365	,581	,3691	-1,5556	,1,9116
V6 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,169	,882	,1,677	59	,145	,758	,5129	-1,7839	,2687	
				,1,645	,1,111	,297	,768	,4107	-4,8888	,3,3736
V7 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	1,740	,193	,297	54	,693	,483	,1,7473	-2,9834	,1,9982	
				,269	,1,627	,454	,483	,5004	-3,2377	,2,2525
V8 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,329	,867	,079	60	,930	,02	,269	-1,598	,553	
				,067	,2,145	,953	,02	,339	-1,392	,1,346
V9 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	1,391	,186	,593	54	,556	,4,68	,6,893	-17,899	,9,740	
				,1,445	,1,036	,311	,4,68	,2,824	-16,347	,11,188
V10 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,354	,664	,132	62	,096	,070	,6296	-,9886	,1,1205	
				,160	,2,844	,092	,070	,4,392	-1,1950	,1,3050
V11 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	4,437	,041	,544	42	,589	,383	,7029	-1,8001	,1,0349	
				,2421	,34,130	,021	,383	,1,580	-,7037	,0614
V2 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,434	,518	,133	51	,885	,871	,5951	-1,0035	,1,1448	
				,174	,2,470	,075	,871	,4,053	-1,2907	,1,5320
V3 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,010	,919	,149	60	,892	,163	,1,0917	-2,0303	,2,3663	
				,147	,3,500	,892	,163	,1,090	-3,0979	,3,4229

Prueba de varianzas independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						95% Intervalo de confianza para la diferencia	
	F	Sig.	t	gr	Sig. (bilateral)	Diferencia de medias	Error std. de la diferencia			
								inferior	superior	
V8 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,079	,780	,017	59	,313	,315	,3096	,3047	,9344	
				,1,140	,6,564	,291	,315	,2743	,3426	,9723
V7 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,099	,893	,429	54	,670	,349	,8993	-1,2747	,1,9665	
				,376	,4,664	,724	,348	,9200	-2,0833	,2,7760
V8 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,069	,817	,173	60	,483	,16	,208	,255	,577	
				,091	,5,845	,516	,16	,233	,412	,723
V9 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,326	,466	,695	54	,496	,3,11	,4,466	-12,095	,6,671	
				,550	,4,461	,808	,3,11	,5,653	-18,188	,11,985
V10 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,029	,965	,663	62	,510	,271	,4,093	-1,0993	,5,669	
				,701	,7,806	,603	,271	,3,866	-1,1686	,8242
V1 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,210	,575	,097	49	,931	,045	,5107	-1,0744	,9054	
				,096	,3,659	,935	,045	,4,956	-1,4726	,1,3835
V2 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	2,639	,111	,241	51	,811	,094	,3,901	,6,692	,8771	
				,307	,10,423	,706	,094	,3,425	,4,435	,6,314
V3 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	,561	,457	,050	50	,961	,054	,1,0919	-2,1380	,2,2474	
				,044	,3,399	,967	,054	,1,2187	-3,5787	,3,6871
V4 Se han assumido varianzas iguales. No se han assumido varianzas iguales.	2,752	,183	,142	58	,888	,088	,646	,3,372	,1,195	
				,295	,14,941	,779	,088	,210	,769	,599

Se observa que para prácticamente todas las variables, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de cada una de ellas (los intervalos de confianza contienen el valor cero). Por lo tanto se puede concluir con bastante fiabilidad la distribución aleatoria de los datos perdidos, conclusión que permitirá realizar análisis estadísticos con los datos aplicando distintos métodos de imputación de la información faltante.

Antes de realizar cualquier análisis de datos ausentes es conveniente realizar un primer análisis descriptivo de la información que presente la distribución de los valores perdidos por variables y otras medidas de centralización y dispersión. Esta tarea puede llevarse a cabo, por ejemplo, con el procedimiento *Frecuencias* de SPSS. Para ello se elige *Analizar → Estadísticos descriptivos → Frecuencias* (Figura 3-121) y se introducen las variables en la pantalla *Frecuencia* (Figura 3-122). El botón *Estadísticos* permite elegir los estadísticos a presentar. Al pulsar *Aceptar* se obtiene ya el informe titulado *Estadísticos*.

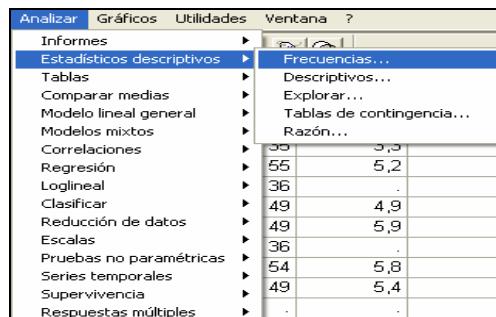


Figura 3-121

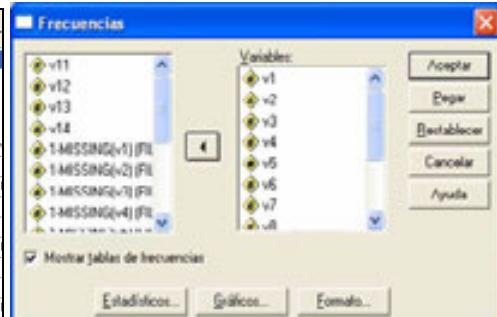


Figura 3-122

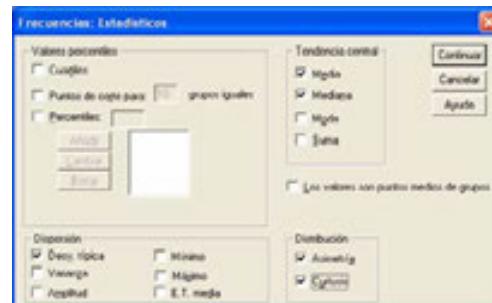


Figura 3-123

		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
N	Válidos	45	53	52	60	57	61	56	62	56	54
	Perdidos	19	11	12	4	7	3	8	2	0	0
Media		6,016	1,900	7,760	6,06	2,911	2,833	6,826	,36	46,43	4,684
Mediana		3,800	1,800	8,200	5,95	3,100	2,800	6,800	,00	47,00	4,800
Desv. Sp.		,3638	,0816	,0076	,475	,7624	,7204	,7130	,483	9,516	1,0173
Alometría		,651	,299	,7,068	,1,245	,379	,330	,404	,327	,295	,5,525
Error Sp. de asimetría		,354	,327	,300	,308	,316	,308	,304	,319	,299	
Curvosis		,387	,108	6,787	,3,176	,098	,311	,092	1,668	,575	5,682
Prueba de curvosis		,695	,644	,650	,608	,623	,604	,628	,599	,628	,690

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico con ellos. Podemos comenzar incluyendo sólo en el análisis las observaciones (casos) con datos completos (filas cuyos valores para todas las variables sean válidos), es decir, cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se denomina *aproximación de casos completos* o *supresión de casos según lista* y suele ser el método por defecto en la mayoría del software estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, porque su supresión provocaría una muestra representativa de la información total. En caso contrario se reduciría mucho el tamaño de la muestra a considerar para el análisis y no sería representativa de la información completa. Otro método consiste en la *supresión de datos según pareja*, es decir, se trabaja con todos los casos (filas) posibles que tengan valores válidos para cada par de variables que se consideren en el análisis independientemente de lo que ocurra en el resto de las variables. Este método elimina menos información y se utiliza siempre en cualquier análisis bivariante o transformable en bivariante. Otro método adicional consiste en *suprimir los casos (filas) o variables (columnas)* que peor se comportan respecto a los datos ausentes. Nuevamente es necesario sopesar la cantidad de datos a eliminar. Debe siempre considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable o conjunto de casos en el análisis estadístico.

La alternativa a los métodos de supresión de datos es la *imputación de la información faltante*. La imputación es el proceso de estimación de valores ausentes basado en valores válidos de otras variables o casos de la muestra. Para realizar la imputación con SPSS elija en los menús: *Transformar → Reemplazar valores perdidos* (Figura 3-124) y seleccione el método de estimación que deseé utilizar para reemplazar los valores perdidos (Figura 3-125). Seleccione la variable o variables para las que desea reemplazar los valores perdidos. Se pueden imputar los valores perdidos sustituyéndolos por la media de la serie, por la media o mediana de los puntos adyacentes, por interpolación lineal o por tendencia lineal en el punto. Se utilizará imputación por la media. Al pulsar *Aceptar* se obtiene un informe previo sobre las variables imputadas y SPSS incorpora al editor de datos las nuevas variables imputadas con nuevos nombres para no perder las antiguas sin imputar (Figura 3-126).

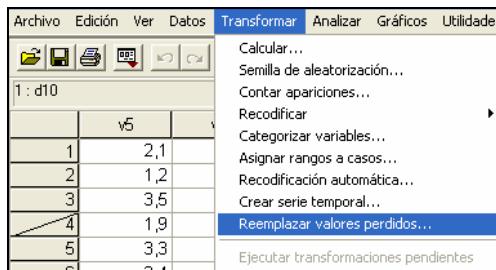


Figura 3-124

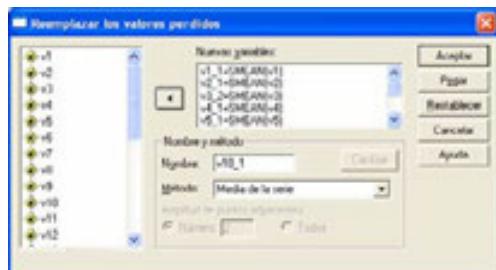


Figura 3-125

Result Variable	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
V1_1	21	1	70	70	SMEAN (V1)
V2_1	12	1	70	70	SMEAN (V2)
V3_2	15	1	70	70	SMEAN (V3)
V4_1	5	1	70	70	SMEAN (V4)
V5_1	9	1	70	70	SMEAN (V5)
V6_1	6	1	70	70	SMEAN (V6)
V7_1	9	1	70	70	SMEAN (V7)
V8_1	3	1	70	70	SMEAN (V8)
V9_1	9	1	70	70	SMEAN (V9)
V10_1	6	1	70	70	SMEAN (V10)

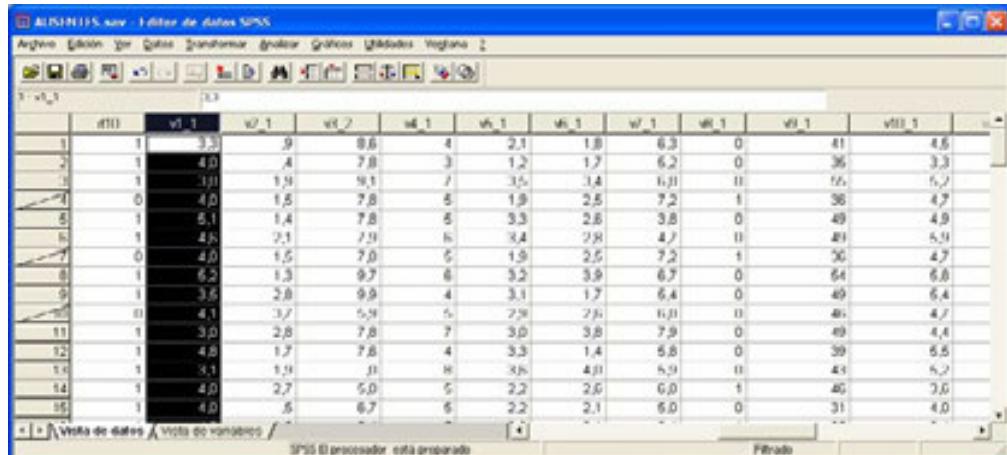


Figura 3-126

Para el *análisis de los datos atípicos* se puede utilizar el diagrama de caja y bigotes. Para ello, puede utilizarse *Gráficos* → *Diagramas de caja* (Figura 3-127) y elegir *Simple* y *Resúmenes para distintas variables* en la Figura 3-128. Se hace clic en *Definir* y en el campo *Las cajas representan* de la Figura 3-129 introducimos todas las variables (las imputadas) para las que queremos el gráfico de caja y bigotes. El botón *Opciones* ya no tiene sentido, porque el tratamiento previo de los datos ausentes nos ha llevado a la imputación de la información faltante, con lo que para las variables imputadas ya no hay valores ausentes. Al pulsar en *Aceptar* se obtiene el gráfico de caja y bigotes para cada una de las variables imputadas de la Figura 3-130. Según este gráfico hay valores atípicos en casi todas las variables, algunos de los cuales son extremos (los tachados). Además, todos los valores atípicos reflejan sobre el gráfico el valor de su dato correspondiente. No se ha utilizado la variable V9 imputada porque sus valores tienen una escala muy distinta. Puede analizarse aparte con un gráfico de caja y bigotes para ella sola y se observa que no presenta valores atípicos.



Figura 3-127



Figura 3-128



Figura 3-129

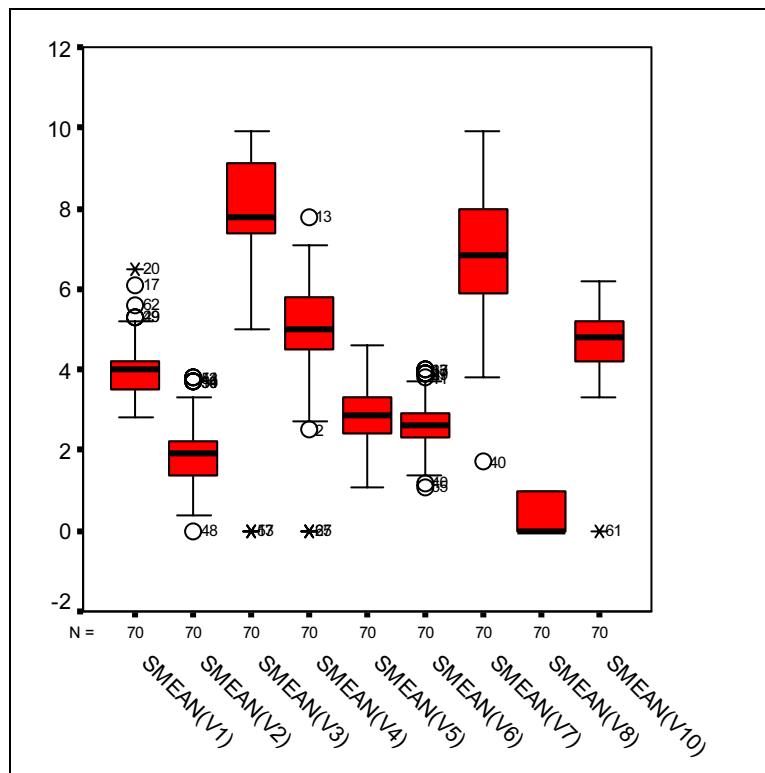


Figura 3-130

Las variables V6, V1, V2 y V4 imputadas presentan valores atípicos dudosos en el gráfico de cajas. Para aclaración adicional se realizarán gráficos de control para ellas mediante *Gráficos → Control → Individuos, rango móvil* (Figura 3-131) y rellenando el campo *Medida del proceso* de la Figura 3-132 con la variable V6 imputada. Al hacer clic en *Aceptar* se obtiene el gráfico de la Figura 3-133 que no presenta ningún valor atípico claro que sobrepase las líneas de control para V6. Los Gráficos 3-134 a 3-136 presentan los gráficos de control para las otras variables.

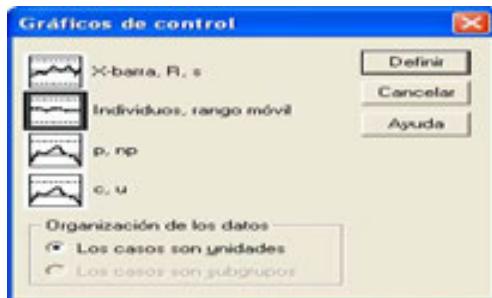


Figura 3-131

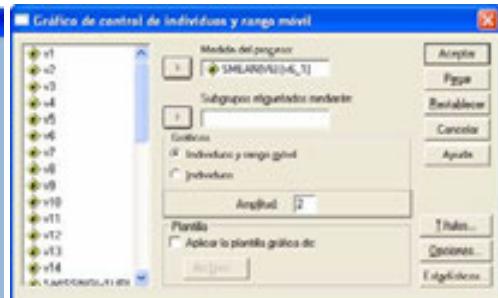


Figura 3-132

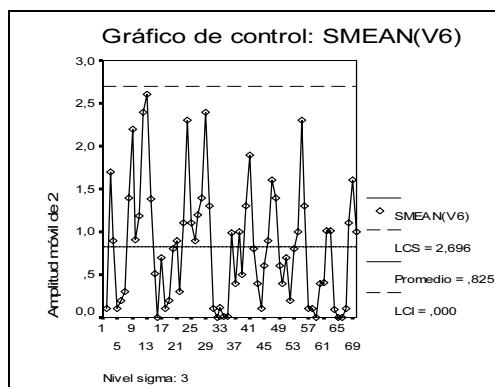


Figura 3-133

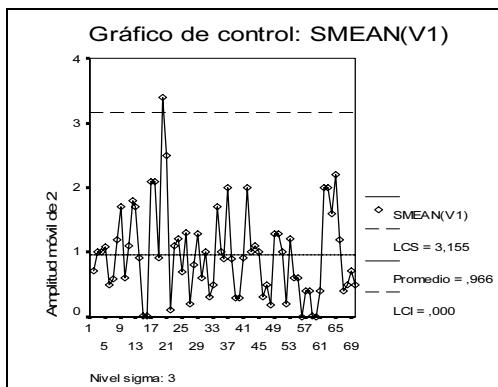


Figura 3-134

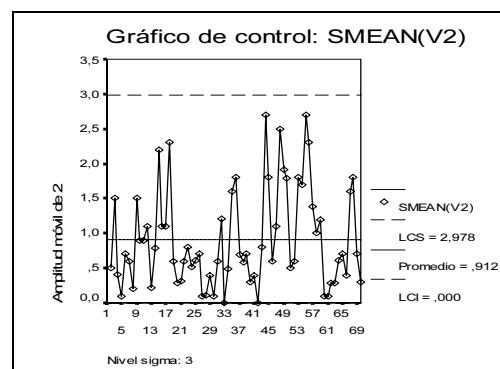


Figura 3-135

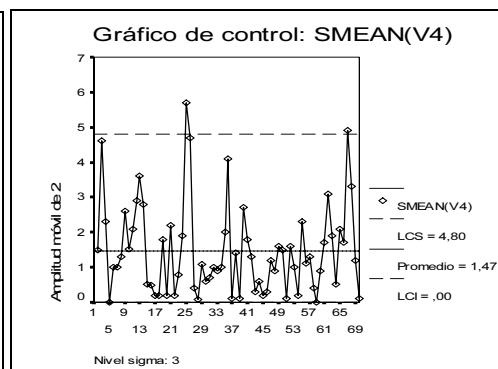


Figura 3-136

ANÁLISIS EN COMPONENTES PRINCIPALES

OBJETIVO DEL ANÁLISIS EN COMPONENTES PRINCIPALES

El análisis en componentes principales es una técnica de análisis estadístico multivariante que se clasifica entre los métodos de simplificación o reducción de la dimensión y que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos persiguiendo obtener un menor número de variables, combinación lineal de las primitivas, que se denominan componentes principales o factores, cuya posterior interpretación permitirá un análisis más simple del problema estudiado. Su aplicación es directa sobre cualquier conjunto de variables, a las que considera en bloque, sin que el investigador haya previamente establecido jerarquías entre ellas, ni necesite comprobar la normalidad de su distribución. Se trata por tanto de una técnica para el análisis de la interdependencia (en contraposición con las técnicas de la dependencia).

El análisis en componentes principales permite describir, de un modo sintético, la estructura y las interrelaciones de las variables originales en el fenómeno que se estudia a partir de las componentes obtenidas que, naturalmente, habrá que interpretar y «nombrar». El mayor número posible de componentes coincide, como veremos, con el número total de variables. Quedarse con todas ellas no simplificaría el problema, por lo que el investigador deberá seleccionar entre distintas alternativas aquéllas que, siendo pocas e interpretables, expliquen una proporción aceptable de la varianza global o inercia de la nube de puntos que suponga una razonable pérdida de información. Esta reducción de muchas variables a pocas componentes puede simplificar la aplicación sobre estas últimas de otras técnicas multivariantes (regresión, clusters, etc.).

El método de componentes principales tiene por objeto transformar un conjunto de variables, a las que denominaremos variables *originales interrelacionadas*, en un nuevo conjunto de variables, combinación lineal de las originales, denominadas *componentes principales*. Estas últimas se caracterizan por estar incorrelacionadas entre sí.

En cuanto al interés que tiene esta técnica, en muchas ocasiones el investigador se enfrenta a situaciones en las que, para analizar un fenómeno, dispone de información de muchas variables que están correlacionadas entre sí en mayor o menor grado. Estas correlaciones son como un velo que impiden evaluar adecuadamente el papel que juega cada variable en el fenómeno estudiado. El análisis de componentes principales permite pasar a un nuevo conjunto de variables, las componentes principales, que gozan de la ventaja de estar incorrelacionadas entre sí y que, además, pueden ordenarse de acuerdo con la información que llevan incorporada. Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada dicha componente. Por esta razón se selecciona como primera componente aquélla que tenga mayor varianza, mientras que, por el contrario, la última es la de menor varianza.

En general, la extracción de componentes principales se efectúa sobre variables *tipificadas* para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en *desviaciones* respecto a la media. Si p variables están tipificadas, la suma de las varianzas es igual a p , ya que la varianza de una variable tipificada es por definición igual a 1. El nuevo conjunto de variables que se obtienen por el método de componentes principales, es igual en número al de variables originales. Es importante destacar que la suma de sus varianzas es igual a la suma de las varianzas de las variables originales. Las diferencias entre ambos conjuntos de variables estriba en que, como ya se ha indicado, las componentes principales se calculan de forma que estén incorrelacionadas entre sí. Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes. Si las variables originales estuvieran completamente incorrelacionadas entre sí, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirían con las variables originales.

Merece hacer hincapié en que las componentes principales se expresan como una combinación lineal de las variables originales. Desde el punto de vista de su aplicación, el método de componentes principales es considerado como un método de *reducción*, es decir, un método que permite *reducir* la dimensión del número de variables que inicialmente se han considerado en el análisis. Es vital abordar las técnicas usuales para determinar el número de componentes principales a retener. Ésta es una cuestión importante, ya que ese conjunto de componentes retenidas es el que se utilizará en análisis posteriores para representar a todo el conjunto de variables iniciales.

No obstante puede considerarse el método de componentes principales como un método para la reducción de datos, y tratar otros problemas como el de rotación de factores, contrastes, etc. en el método de análisis factorial que implica una mayor formalización. En este sentido, el método de componentes principales se inscribe dentro de la estadística descriptiva.

OBTENCIÓN DE LAS COMPONENTES PRINCIPALES

En el análisis en componentes principales se dispone de una muestra de tamaño n acerca de p variables X_1, X_2, \dots, X_p (tipificadas o expresadas en desviaciones respecto de su media) inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número $k \leq p$ de variables incorrelacionadas Z_1, Z_2, \dots, Z_p que sean combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad.

La primera componente principal, al igual que las restantes, se expresa como combinación lineal de las variables originales como sigue:

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \dots + u_{1p}X_{pi}$$

Para el conjunto de las n observaciones muestrales esta ecuación puede expresarse matricialmente como sigue:

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & & & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

En notación abreviada tendremos: $Z_1 = X u_1$

Tanto si las X_j están tipificadas, como si están expresadas en desviaciones respecto de su media muestral, la media de Z_1 es cero, esto es, $E(Z_1) = E(X u_1) = E(X)u_1 = 0$.

La varianza de Z_1 será:

$$V(Z_1) = \frac{\sum_{i=1}^n Z_{1i}^2}{n} = \frac{1}{n} Z_1' Z_1 = \frac{1}{n} u_1' X' X u_1 = u_1' \left[\frac{1}{n} X' X \right] u_1 = u_1' V u_1$$

Si las variables están expresadas en desviaciones respecto a la media, la expresión $\frac{1}{n}X'X$ (**matriz de inercia**) es la matriz de covarianzas muestral a la que denominaremos V (caso más general) y para variables tipificadas $\frac{1}{n}X'X$ es la matriz de correlaciones R .

La primera componente Z_1 se obtiene de forma que su varianza sea máxima y sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, es decir, la variable de los pesos o ponderaciones $(u_{11}, u_{12}, \dots, u_{1p})'$ se toma normalizada.

Se trata entonces de hallar Z_1 maximizando $V(Z_1) = u_1'Vu_1$, sujeta a la restricción $\sum_{j=1}^p u_{1j}^2 = u_1'u_1 = 1$

Para resolver este problema de optimización con restricciones se aplica el método de los multiplicadores de Lagrange considerando la función lagrangiana:

$$L = u_1'Vu_1 - \lambda(u_1'u_1 - 1)$$

Derivando respecto de u_1 e igualando a cero, se tiene:

$$\frac{\partial L}{\partial u_1} = 2Vu_1 - 2\lambda u_1 = 0 \Rightarrow (V - \lambda I)u_1 = 0$$

Se trata de un sistema homogéneo en u_1 , que sólo tiene solución si el determinante de la matriz de los coeficientes es nulo, es decir, $|V - \lambda I| = 0$. Pero la expresión $|V - \lambda I| = 0$ es equivalente a decir que λ es un valor propio de la matriz V .

En general, la ecuación $|V - \lambda I| = 0$ tiene n raíces $\lambda_1, \lambda_2, \dots, \lambda_n$, que puedo ordenarlas de mayor a menor $\lambda_1 > \lambda_2 > \dots > \lambda_n$.

En la ecuación $(V - \lambda I)u_1 = 0$ podemos multiplicar por u_1' a la derecha, con lo que se tiene $u_1'(V - \lambda I)u_1 = 0 \Rightarrow u_1'Vu_1 = \lambda \Rightarrow V(Z_1) = \lambda$. Por lo tanto, para maximizar $V(Z_1)$ he de tomar el mayor valor propio λ de la matriz V .

Tomando λ_1 como el mayor valor propio de V y tomando u_1 como su vector propio asociado normalizado ($u_1'u_1 = 1$), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la primera componente principal, componente que vendrá definida como:

$$Z_1 = X u_1$$

La segunda componente principal, al igual que las restantes, se expresa como combinación lineal de las variables originales como sigue:

$$Z_{2i} = u_{21}X_{1i} + u_{22}X_{2i} + \cdots + u_{2p}X_{pi}$$

Para el conjunto de las n observaciones muestrales esta ecuación puede expresarse matricialmente como sigue:

$$\begin{bmatrix} Z_{21} \\ Z_{22} \\ \vdots \\ Z_{2n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ & & \vdots & \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2p} \end{bmatrix}$$

En notación abreviada tendremos: $Z_2 = X u_2$

Tanto si las X_j están tipificadas, como si están expresadas en desviaciones respecto de su media muestral, la media de Z_2 es cero, esto es, $E(Z_2) = E(X u_2) = E(X)u_2 = 0$.

La varianza de Z_2 será:

$$V(Z_2) = \frac{\sum_{i=1}^n Z_{2i}^2}{n} = \frac{1}{n} Z_2' Z_2 = \frac{1}{n} u_2' X' X u_2 = u_2' \left[\frac{1}{n} X' X \right] u_2 = u_2' V u_2$$

La segunda componente Z_2 se obtiene de forma que su varianza sea máxima sujeta a la restricción de que la suma de los pesos u_{2j} al cuadrado sea igual a la unidad, es decir, la variable de los pesos o ponderaciones $(u_{21}, u_{22}, \dots, u_{2p})'$ se toma normalizada ($u_2' u_2 = 1$).

Por otra parte como Z_1 y Z_2 han de estar incorrelacionados se tiene que:

$$0 = E(Z_2' Z_1) = E(u_2' X' X u_1) = u_2' E(X' X) u_1 = u_2' V u_1$$

También sabemos que $V u_1 = \lambda_1 u_1$ (ya que u_1 es el vector propio de V asociado a su mayor valor propio λ_1). Si multiplicamos por u_2' a la derecha tenemos:

$$0 = u_2' V u_1 = \lambda_1 u_2' u_1 \Rightarrow u_2' u_1 = 0$$

con lo que u_2 y u_1 son ortogonales.

Se trata entonces de hallar Z_2 maximizando $V(Z_2) = u_2'Vu_2$, sujeta a las restricciones $u_2'u_2=1$ y $u_2'Vu_1=0$.

Para resolver este problema de optimización con dos restricciones se aplica el método de los multiplicadores de Lagrange considerando la función lagrangiana:

$$L = u_2'Vu_2 - 2\mu(u_2'Vu_1) - \lambda(u_2'u_2 - 1)$$

Derivando respecto de u_2 e igualando a cero, se tiene:

$$\frac{\partial L}{\partial u_2} = 2Vu_2 - 2\mu Vu_1 - 2\lambda u_2 = 0$$

Dividiendo por 2 y premultiplicando por u_1' tenemos:

$$u_1'Vu_2 - \mu u_1'Vu_1 - \lambda u_1'u_2 = 0$$

y como $Vu_1 = \lambda_1 u_1$ (ya que u_1 es el vector propio de V asociado a su mayor valor propio λ_1), entonces $u_1'V = \lambda_1 u_1'$, y podemos escribir la igualdad anterior como:

$$\lambda_1 u_1'u_2 - \mu V[Z_1] - \lambda u_1'u_2 = 0$$

Pero:

$$u_1'u_2 = 0 \Rightarrow \mu V[Z_1] = 0 \Rightarrow \mu = 0$$

De donde:

$$\frac{\partial L}{\partial u_2} = 2Vu_2 - 2\lambda u_2 = 0 \Rightarrow (V - \lambda I)u_2 = 0$$

Se trata de un sistema homogéneo en u_2 , que sólo tiene solución si el determinante de la matriz de los coeficientes es nulo, es decir, $|V - \lambda I| = 0$. Pero la expresión $|V - \lambda I| = 0$ es equivalente a decir que λ es un valor propio de la matriz V .

En general, la ecuación $|V - \lambda I| = 0$ tiene n raíces $\lambda_1, \lambda_2, \dots, \lambda_n$, que puedo ordenarlas de mayor a menor $\lambda_1 > \lambda_2 > \dots > \lambda_n$.

En la ecuación $(V - \lambda I)u_2 = 0$ podemos multiplicar por u_2' a la derecha, con lo que se tiene $u_2'(V - \lambda I)u_2 = 0 \Rightarrow u_2'Vu_2 = \lambda \Rightarrow V(Z_2) = \lambda$. Por lo tanto, para maximizar $V(Z_2)$ he de tomar el segundo mayor valor propio λ de la matriz V (el mayor ya lo había tomado al obtener la primera componente principal).

Tomando λ_2 como el segundo mayor valor propio de V y tomando u_2 como su vector propio asociado normalizado ($u_2'u_2=1$), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la segunda componente principal, componente que vendrá definida como:

$$Z_2 = X u_2$$

De forma similar, **la componente principal h-ésima** se define como $Z_h = Xu_h$ donde u_h es el vector propio de V asociado a su h-ésimo mayor valor propio. Suele denominarse también a u_h **eje factorial h-ésimo**.

VARIANZAS DE LAS COMPONENTES

En el proceso de obtención de las componentes principales presentado en el apartado anterior hemos visto que la varianza de la componente h-ésima es:

$$V(Z_h) = u'_h V u_h = \lambda_h$$

Es decir, la varianza de cada componente es igual al valor propio de la matriz V al que va asociada.

Si, como es lógico, la medida de la variabilidad de las variables originales es la suma de sus varianzas, dicha variabilidad será:

$$\sum_{h=1}^p V(X_h) = \text{traza}(V)$$

ya que las varianzas de las variables son los términos que aparecen en la diagonal de la matriz de varianzas covarianzas V .

Ahora bien, como V es una matriz real simétrica, por la teoría de diagonalización de matrices, existe una matriz ortogonal P ($P^{-1}=P'$) tal que $P'VP=D$, siendo D diagonal con los valores propios de V ordenados de mayor a menor en la diagonal principal. Por lo tanto:

$$\text{traza}(P'VP) = \text{traza}(D) = \sum_{h=1}^p \lambda_h$$

Pero:

$$\text{traza}(P'VP) = \text{traza}(VPP') = \text{traza}(V.I) = \text{traza}(V)$$

Con lo que ya podemos escribir:

$$\sum_{h=1}^p V(X_h) = \text{traza}(V) = \text{traza}(P'VP) = \text{traza}(D) = \sum_{h=1}^p \lambda_h = \sum_{h=1}^p V(Z_h)$$

Hemos comprobado, además, que la suma de las varianzas de las variables (**inerzia total de la nube de puntos**) es igual a la suma de las varianzas de las componentes principales e igual a la suma de los valores propios de la matriz de varianzas covarianzas muestral V .

La proporción de la variabilidad total recogida por la componente principal h-ésima (*porcentaje de inercia explicada por la componente principal h-ésima*) vendrá dada por:

$$\frac{\lambda_h}{\sum_{h=1}^p \lambda_h} = \frac{\lambda_h}{\text{traza}(V)}$$

Si las variables están tipificadas, $V = R$ y $\text{traza}(V) = \text{traza}(R) = p$, con lo que la proporción de la componente h-ésima en la variabilidad total será λ_h/p .

También se define el *porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales)* como:

$$\frac{\sum_{h=1}^k \lambda_h}{\sum_{h=1}^p \lambda_h} = \frac{\sum_{h=1}^k \lambda_h}{\text{traza}(V)}$$

ESTRUCTURA FACTORIAL DE LAS COMPONENTES PRINCIPALES

Se denomina estructura factorial de las componentes principales a la matriz de correlaciones entre las componentes Z_h y las variables originales X_j .

Consideramos los vectores muestrales relativos a Z_h y X_j respectivamente:

$$X_j = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jn} \end{bmatrix} \quad Z_h = \begin{bmatrix} Z_{h1} \\ Z_{h2} \\ \vdots \\ Z_{hn} \end{bmatrix}$$

La covarianza muestral entre Z_h y X_j viene dada por:

$$\text{Cov}(X_j, Z_h) = \frac{1}{n} X'_j Z_h$$

El vector X_j se puede expresar en función de la matriz X utilizando el vector de orden p , al que denominamos por δ , que tiene un 1 en la posición j-ésima y 0 en las posiciones restantes. La forma de expresar X_j en función de la matriz X a través del vector δ es la siguiente:

$$X'_j = \delta' X' = [0 \ \cdots \ 1 \ \cdots \ 0] \begin{bmatrix} X_{11} & \cdots & X_{1i} & \cdots & X_{1n} \\ \vdots & & \vdots & & \vdots \\ X_{j1} & \cdots & X_{ji} & \cdots & X_{jn} \\ \vdots & & \vdots & & \vdots \\ X_{p1} & \cdots & X_{pi} & \cdots & X_{pn} \end{bmatrix}$$

Teniendo en cuenta que $Z_h = X u_h$ podemos escribir:

$$Cov(X_j, Z_h) = \frac{1}{n} X'_j Z_h = \frac{1}{n} \delta' X' X u_h = \delta' V u_h = \delta' \lambda_h u_h = \lambda_h \delta' u_h = \lambda_h u_{hj}$$

Por lo tanto, podemos escribir la correlación existente entre la variable X_j y la componente Z_h de la siguiente forma:

$$r_{jh} = \frac{Cov(X_j, Z_h)}{\sqrt{V(X_j)} \sqrt{V(Z_h)}} = \frac{\lambda_h u_{hj}}{\sqrt{V(X_j)} \sqrt{\lambda_h}}$$

Si las variables originales están tipificadas, la correlación entre las variable X_j y la componente Z_h es la siguiente:

$$r_{jh} = \frac{\lambda_h u_{hj}}{\sqrt{V(X_j)} \sqrt{\lambda_h}} = \frac{\lambda_h u_{hj}}{\sqrt{\lambda_h}} = u_{hj} \sqrt{\lambda_h}$$

PUNTUACIONES O MEDICIÓN DE LAS COMPONENTES

El análisis en componentes principales es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por las componentes obtenidas. Por ejemplo, en el caso de estimación de modelos afectados de multicolinealidad o correlación serial (autocorrelación). Por ello, es necesario conocer los valores que toman las componentes en cada observación.

Una vez calculados los coeficientes u_{hj} (componentes del vector propio normalizado asociado al valor propio h -ésimo de la matriz $V = X'X/n$ relativo a la componente principal Z_h), se pueden obtener las puntuaciones Z_{hj} , es decir, los valores de las componentes correspondientes a cada observación, a partir de la siguiente relación:

$$Z_{hi} = u_{h1} X_{1i} + u_{h2} X_{2i} + \cdots + u_{hp} X_{pi} \quad h = 1 \dots p \quad i = 1 \dots n$$

Si las componentes se dividen por su desviación típica se obtienen las componentes tipificadas. Por lo tanto, si llamamos Y_h a la componente Z_h tipificada tenemos:

$$Y_h = \frac{Z_h - E(Z_h)}{\sqrt{V(Z_h)}} = \frac{Z_h}{\sqrt{\lambda_h}} \quad h = 1 \dots p$$

Por lo tanto, las puntuaciones tipificadas serán:

$$\frac{Z_{hi}}{\sqrt{\lambda_h}} = \frac{u_{h1}}{\sqrt{\lambda_h}} X_{1i} + \frac{u_{h2}}{\sqrt{\lambda_h}} X_{2i} + \dots + \frac{u_{hp}}{\sqrt{\lambda_h}} X_{pi} \quad h = 1 \dots p \quad i = 1 \dots n$$

expresión que puede escribirse como:

$$Y_{hi} = c_{h1} X_{1i} + c_{h2} X_{2i} + \dots + c_{hp} X_{pi} \quad c_{hi} = \frac{u_{hi}}{\sqrt{\lambda_h}} \quad h = 1 \dots p \quad i = 1 \dots n$$

La matriz formada por los coeficientes c_{hi} suele denominarse matriz de coeficientes de puntuaciones de los factores (*factor score coefficient matrix*).

CONTRASTES SOBRE EL NÚMERO DE COMPONENTES PRINCIPALES A RETENER

En general, el objetivo de la aplicación de las componentes principales es reducir las dimensiones de las variables originales, pasando de p variables originales a $m < p$ componentes principales. El problema que se plantea es cómo fijar m , o, dicho de otra forma, ¿qué número de componentes se deben retener? Aunque para la extracción de las componentes principales no hace falta plantear un modelo estadístico previo, algunos de los criterios para determinar cuál debe ser el número óptimo de componentes a retener requieren la formulación previa de hipótesis estadísticas.

Criterio de la media aritmética

Según este criterio se seleccionan aquellas componentes cuya raíz característica λ_j excede de la media de las raíces características. Recordemos que la raíz característica asociada a una componente es precisamente su varianza.

Analíticamente este criterio implica retener todas aquellas componentes en que se verifique que:

$$\lambda_h > \bar{\lambda} = \frac{\sum_{j=1}^p \lambda_j}{p}$$

Si se utilizan variables tipificadas, entonces, como ya se ha visto, se verifica que $\sum_{j=1}^p \lambda_j = p$, con lo que para variables tipificadas se retiene aquellas componentes tales que $\lambda_j > 1$.

Contraste sobre las raíces características no retenidas

Se puede considerar que, las $p-m$ últimas raíces características poblacionales son iguales a 0. Si las raíces muestrales que observamos correspondientes a estas componentes no son exactamente igual a 0, se debe a los problemas del azar. Por ello, bajo el supuesto de que las variables originales siguen una distribución normal multivariante, se pueden formular las siguientes hipótesis relativas a las raíces características poblacionales:

$$H_0: \lambda_{m+1} = \lambda_{m+2} = \dots = \lambda_p = 0$$

El estadístico que se considera para contrastar esta hipótesis es el siguiente:

$$Q^* = \left(n - \frac{2p+11}{6} \right) \left((p-m) \ln \bar{\lambda}_{p-m} - \sum_{j=m+1}^p \ln \lambda_j \right)$$

Bajo la hipótesis nula H_0 , el estadístico anterior se distribuye como una chi-cuadrado con $(p-m+2)(p-m+1)/2$ grados de libertad. Este contraste se deriva del contraste de esfericidad de Barlett para la existencia o no de una relación significativa entre las variables analizadas que se utiliza en la validación del modelo de análisis multivariante de la varianza.

Para ver la mecánica de la aplicación de este contraste, supongamos que inicialmente se han retenido m raíces características (por ejemplo, las que superan la unidad) al aplicar el criterio de la media aritmética. En el caso de que se rechace la hipótesis nula H_0 , implica que una o más de las raíces características no retenidas es significativa. La decisión a tomar en ese caso sería retener una nueva componente, y aplicar de nuevo el contraste a las restantes raíces características. Este proceso continuaría hasta que no se rechace la hipótesis nula.

Prueba de Anderson

Si los valores propios, a partir del valor $m+1$, son iguales, no hay ejes principales a partir del eje $m+1$, en el sentido de que no hay direcciones de máxima variabilidad. La variabilidad en las últimas $(n-m)$ dimensiones es esférica. Para decidir este hecho se debe testearse la hipótesis siguiente:

$$H_0: \lambda_{m+1} = \lambda_{m+2} = \dots = \lambda_p$$

Si esta hipótesis es cierta, el estadístico:

$$\chi^2 = -(n-1) \sum_{j=m+1}^p \ln \lambda_j + (p-m)(n-1) \ln \left(\frac{\sum_{j=m+1}^p \ln \lambda_j}{(p-m)} \right)$$

sigue una distribución chi-cuadrado con $(p-m)(p-m+1)/2-1$ grados de libertad, siempre y cuando el número de individuos n sea grande. Si para un m fijado, χ^2 es significativo, debe rechazarse la hipótesis H_0 . $\lambda_1, \dots, \lambda_n$ representan los valores propios calculados sobre la matriz de covarianzas muestral.

Esta prueba sólo es válida si las variables X_1, \dots, X_n son normales con distribución conjunta normal.

Prueba de Lebart y Fenelón

Tanto esta prueba como las dos siguientes obedecen a una concepción más empírica que racional del problema. La formulación matemática de lo que pretenden demostrar está pobemente justificada en términos de inferencia estadística.

La idea general es la siguiente: a partir de una cierta dimensión (número de componentes a retener), la restante variabilidad explicada es debida a causas aleatorias (ruidos) que perturban la información contenida en la tabla de datos inicial. En esencia, este "ruido" es debido a fluctuaciones del muestreo (desviaciones de la normalidad, errores de medida, gradientes de dependencia entre los individuos, etc.). Asimilando el ruido a variables independientes, la significación de la dimensión m queda resuelta cuando la varianza explicada supera claramente a la varianza explicada por el ruido. La varianza explicada por las primeras m componentes viene expresada por $V_m = \lambda_1 + \dots + \lambda_m$.

La prueba de Lebart y Fenelon consiste en realizar k análisis sobre n variables independientes para un tamaño muestral n . Ordenando las varianzas explicadas en cada análisis tenemos que $V_m^{i_1} < V_m^{i_2} \dots < V_m^{i_k}$.

La probabilidad de que se verifique una ordenación fijada es $1/k!$. Consideremos el suceso: "la varianza explicada por el k -ésimo análisis supera a la varianza de los demás", es decir, $V_m^{i_1} < V_m^{i_2} \dots < V_m^{i_{k-1}} < V_m^k$. Como podemos formar $(k-1)!$ permutaciones en el conjunto $(1, \dots, k-1)$, la probabilidad de este suceso vendrá dada por $(k-1)! / k! = 1/k$.

Consideremos entonces el nivel de significación $\alpha = 0.05$. Sea V_m la varianza explicada por el análisis real cuya dimensión queremos estudiar. Generemos $k-1 = 19$ ($1/k = 0,05 \Rightarrow k = 100/5 = 20$) análisis con variables independientes generadas al azar. Si V_m procede de variables independientes, la probabilidad de que supere a las varianzas explicadas por los análisis simulados es $1/20 = 0.05$. De este modo tenemos una prueba no paramétrica para decidir la significación de V_m al nivel $\alpha = 0,05$. Si V_m supera a la varianza explicada por los 19 análisis simulados, se puede afirmar, con probabilidad de error 0,05, que la dimensión m es significativa en el sentido dado anteriormente. De manera análoga, para un nivel de significación 0,01 deberíamos simular $k-1 = 99$ análisis ($1/k = 0,01 \Rightarrow k = 100/1 = 100$).

El valor crítico de V_m , a partir del cual la varianza explicada es significativa, se obtiene por simulación de datos generados al azar. Lebart y Fenelon publican gráficas y tablas de V_m para $1 \leq m \leq 5$ en función del número de observaciones n y el número de variables p .

Prueba del bastón roto de Frontier

Frontier asimila la descomposición de la variabilidad total $VT = \lambda_1 + \dots + \lambda_p$ al romper un bastón de longitud VT en p trozos por $p-1$ lugares del mismo elegidos al azar. Ordenando los trozos del bastón, de longitudes $L_1 \geq \dots \geq L_p$, se demuestra que:

$$E(L_p) = \frac{1}{p^2}, \quad E(L_{n-1}) = \frac{1}{p} \left(\frac{1}{p} + \frac{1}{p-1} \right), \quad E(L_j) = \frac{1}{p} \sum_{i=0}^{p-j} \frac{1}{j+1} \quad j = 1, \dots, p$$

Hemos supuesto que $VT = 1$ para normalizar el problema. Si expresamos estos valores medios, cuya suma es 1, en porcentajes de la longitud total, obtenemos el modelo teórico de la descomposición de la varianza en p componentes obtenidas al azar. Por ejemplo, para $p = 4$ tenemos $E(L_1) = 0.5208$, $E(L_2) = 0.2708$, $E(L_3) = 0.1458$ y $E(L_4) = 0.0625$. Por lo que los porcentajes acumulados de varianza de las componentes serán $52,08\%$, $52,08+27,08=79,16\%$, $52,08+27,08+14,58=93,74\%$ y $52,08+27,08+14,58+6,25=100\%$.

Las m primeras componentes son significativas si explican claramente mayor varianza que los m primeros valores medios del modelo del bastón roto. Se considera que las demás componentes descomponen la varianza residual al azar.

Prueba ε de Ibañez

Esta prueba consiste en añadir a las p variables observables del problema una variable ε formada por datos generados al azar. Se repite entonces el análisis de componentes principales con la nueva variable añadida. Si a partir de la componente $m+1$ la variable ε queda resaltada en la estructura factorial (la saturación o carga de ε en la componente $m+1$ es alta), el número significativo de componentes no puede ser superior a m , pues las demás componentes explicarían una variabilidad inferior a la que es debida a la variable arbitraria ε . Ibanez da solamente una justificación empírica de esta prueba, comparando los resultados de un análisis sin variable ε con otro análisis con variable ε , y concluyendo que las componentes deducidas de ambos son prácticamente las mismas. Seguidamente ilustra la prueba ε sobre otros análisis con datos experimentales publicados por el propio Ibanez. La prueba ε sólo llega a proporcionar una cota superior para la dimensión m .

El gráfico de sedimentación

El gráfico de sedimentación se obtiene al representar en ordenadas las raíces características y en abscisas los números de las componentes principales correspondientes a cada raíz característica en orden decreciente. Uniendo todos los puntos se obtiene una Figura que, en general, se parece al perfil de una montaña con una pendiente fuerte hasta llegar a la base, formada por una meseta con una ligera inclinación. Continuando con el símil de la montaña, en esa meseta es donde se acumulan los guijarros caídos desde la cumbre, es decir, donde se sedimentan. Por esta razón, a este gráfico se le conoce con el nombre de gráfico de sedimentación. Su denominación en inglés es *scree plot*. De acuerdo con el criterio gráfico se retienen todas aquellas componentes previas a la zona de sedimentación.

Retención de variables

Hasta ahora todos los contrastes han estado dedicados a determinar el número de componentes a retener. Pero, la retención de componentes, ¿puede afectar a las variables originales? Si se retiene un número determinado de componentes, ¿qué hacer si alguna variable está correlacionada muy débilmente con cada una de las componentes retenidas? Si se plantea un caso de este tipo, sería conveniente suprimir dicha variable del conjunto de variables originales, ya que no estaría representada por las componentes retenidas. Ahora bien, si se considera que la variable a suprimir juega un papel esencial en la investigación, entonces se deberían retener componentes adicionales en el caso de que algunas de ellas estuvieran correlacionadas de forma importante con la variable a suprimir.

LA REGRESIÓN SOBRE COMPONENTES PRINCIPALES Y EL PROBLEMA DE LA MULTICOLINEALIDAD

La regresión sobre componentes principales sustituye el método clásico de ajuste lineal, cuando las variables exógenas del modelo son numerosas o fuertemente correlacionadas entre sí (multicolinealidad).

Consideremos el modelo lineal general $Y = X\beta + e$ con las hipótesis clásicas de normalidad de los residuos, $E(e)=0$ y $V(e) = \sigma^2 I$, pero con problemas de correlación entre las variables exógenas del modelo. Designaremos por \hat{y} el vector de n valores de la variable endógena centrada, y por \hat{X} la matriz conteniendo en columnas los p vectores de n valores, de las variables exógenas centradas. Designaremos estas columnas por $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_p$. Si los vectores $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_p$ no son linealmente independientes (multicolinealidad en el modelo $Y = x\beta + e$), el vector $\hat{\beta} = (\hat{X}' \hat{X})^{-1} \hat{X}' \hat{y}$ de los coeficientes estimados de la regresión no podrá ser calculado, ya que la matriz $\hat{X}' \hat{X}$ no será invertible.

Si algunos de los vectores $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_p$ tienen ángulos pequeños entre sí (dicho de otra forma, si los coeficientes de correlación muestral entre ciertas variables exógenas son cercanos a 1) el vector $\hat{\beta}$ se conocerá, pero con mala precisión. En este caso las contribuciones de cada uno de los coeficientes son difíciles de discernir. En efecto, si la matriz $\hat{X}' \hat{X}$ es «casi singular», algunos de sus valores propios serán próximos a 0. La descomposición de $\hat{X}' \hat{X}$ en función de vectores y valores propios se escribe como:

$$\hat{X}' \hat{X} = \sum_{\alpha=1}^p \lambda_\alpha u_\alpha u'_\alpha$$

ya que $\hat{X}' \hat{X}$ es una matriz simétrica definida positiva con valores propios λ_α relativos a vectores propios u_α ortogonales, cuya diagonalización permite escribir:

$$\hat{X}' \hat{X} = (u_1, \dots, u_p) \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}$$

Además:

$$\hat{X}' \hat{X} = \sum_{\alpha=1}^p \lambda_{\alpha} u_{\alpha} u'_{\alpha} \Rightarrow (\hat{X}' \hat{X})^{-1} = \sum_{\alpha=1}^p \frac{1}{\lambda_{\alpha}} u_{\alpha} u'_{\alpha}$$

La casi nulidad del menor valor propio λ_p de $\hat{X}' \hat{X}$ puede expresarse como:

$$\lambda_p = V(Z_p) = V(\hat{X} u_p) = \frac{1}{n} (\hat{X} u_p)' (\hat{X} u_p) \cong 0 \Rightarrow \hat{X} u_p = 0$$

indicando la casi colinealidad de los vectores columna de \hat{X} . En estas condiciones, el vector de los coeficientes de ajuste mínimo cuadrático se escribe como:

$$\hat{\beta} = (\hat{X}' \hat{X})^{-1} \hat{X}' \hat{y} = \left(\sum_{\alpha=1}^p \frac{1}{\lambda_{\alpha}} u_{\alpha} u'_{\alpha} \right) \hat{X}' \hat{y}$$

y la estimación de su matriz de varianzas covarianzas será:

$$\hat{V}(\hat{\beta}) = S^2 (\hat{X}' \hat{X})^{-1} = S^2 \sum_{\alpha=1}^p \frac{1}{\lambda_{\alpha}} u_{\alpha} u'_{\alpha}$$

lo que permite ver que uno o varios valores propios casi nulos hacen impreciso el ajuste.

Se eliminaría el problema de la casi colinealidad de los vectores columna de \hat{X} suprimiendo $p-q$ vectores u_k ($k = q+1, q+2, \dots, p$) correspondiente a los valores propios λ_k más pequeños de $\hat{X}' \hat{X}$.

En estas condiciones, el vector de los coeficientes de ajuste mínimo cuadrático se escribe como:

$$\hat{\beta}^* = (\hat{X}' \hat{X})^{-1} \hat{X}' \hat{y} = \left(\sum_{\alpha=1}^q \frac{1}{\lambda_{\alpha}} u_{\alpha} u'_{\alpha} \right) \hat{X}' \hat{y} \quad q < p$$

y la estimación de su matriz de varianzas covarianzas será:

$$\hat{V}(\hat{\beta}^*) = S^2 \sum_{\alpha=1}^q \frac{1}{\lambda_{\alpha}} u_{\alpha} u'_{\alpha}$$

Una vez diagonalizada la matriz $\hat{X}'\hat{X}$, el cálculo de los coeficientes de ajuste referidos a (u_1, u_2, \dots, u_q) se realiza considerando las componentes principales tipificadas:

$$z_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \hat{X} u_\alpha \text{ para } \alpha = 1, 2, \dots, q$$

El modelo inicial $Y=X\beta+e$ se ha ajustado ahora mediante $\hat{y}=Zc+d$ donde $Z=(z_1, \dots, z_q)$ es la matriz (n, q) cuyas columnas son los q vectores propios unitarios y ortogonales z_α asociados a los mayores valores propios de $\hat{X}'\hat{X}$, y donde c es el vector de los q nuevos coeficientes hallados mediante:

$$c = (Z'Z)^{-1} \hat{X}' \hat{y} \text{ con } V(c) = S^2(Z'Z)^{-1}$$

Pero como $Z'Z = I_q$ ya que $Z = (z_1, \dots, z_q)$ con z_α ortogonales y unitarios, podemos escribir:

$$c = (Z'Z)^{-1} Z' \hat{y} = Z' \hat{y} \text{ con } V(c) = S^2(Z'Z)^{-1} = S^2 I = \left(\frac{1}{n-q-1} \sum_{i=1}^n d_i^2 \right) I$$

Por lo tanto los coeficiente c están incorrelacionados y tienen todos la misma varianza, estimada por S^2 .

LA REGRESIÓN ORTOGONAL Y LAS COMPONENTES PRINCIPALES

La regresión ortogonal es un método utilizado para determinar una *relación lineal* entre p variables las cuales *a priori* juegan papeles análogos (no se hace la distinción, como en el modelo lineal, entre variables endógenas y exógenas). Más concretamente, se buscan los coeficientes tales que aseguren la *más pequeña dispersión* de esta combinación lineal de las variables.

Sea u un vector de p coeficientes (u_1, \dots, u_p) , sea \hat{X} la matriz (n, p) de observaciones centradas por columnas, y sea $S = \hat{X}' \hat{X} / n$ la matriz de covarianzas muestrales de las p variables. La varianza de la combinación lineal de las variables $Z = \hat{X} u$, definida por u , es la cantidad $V(Z) = V(\hat{X} u) = \frac{1}{n} (\hat{X} u)' (\hat{X} u) = u' S u$.

Bajo este punto de vista, el análisis en componentes principales determina la combinación lineal $Z_1 = \hat{X} u_1$, de u_1 , con *máxima variancia* λ_1 , siendo λ_1 el mayor valor propio de S , y u_1 el vector propio unitario asociado ($u'_1 u_1 = 1$).

$$\lambda_1 = V(Z_1) = V(\hat{X} u_1) = \frac{1}{n} (\hat{X} u_1)' (\hat{X} u_1) = u'_1 S u_1$$

La misma filosofía, aplicada a la búsqueda de la combinación lineal de variables con *varianza mínima*, nos lleva a retener el vector propio u_p de S asociado al más pequeño valor propio λ_p , siendo éste, por otra parte, el valor de esta varianza mínima:

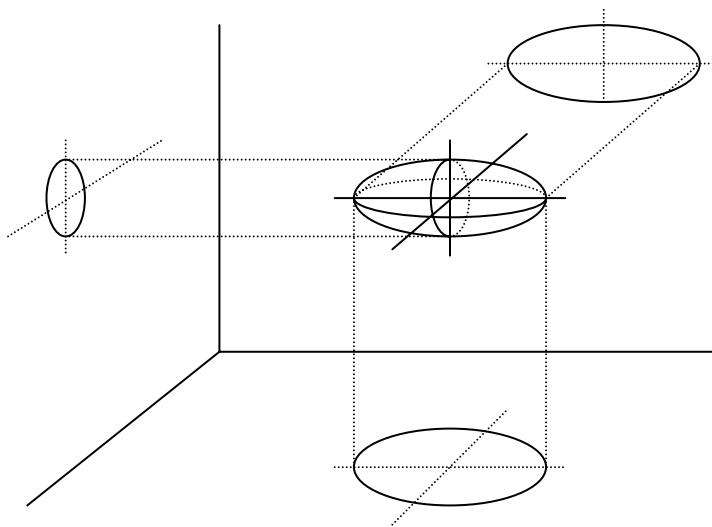
$$\lambda_p = V(Z_p) = V(\hat{X} u_p) = \frac{1}{n} (\hat{X} u_p)' (\hat{X} u_p) = u'_p S u_p$$

Luego, tomando los coeficientes de la regresión ortogonal como las componentes del vector propio u_p de S asociado al más pequeño valor propio λ_p , tenemos caracterizado el mejor ajuste en el sentido de los mínimos cuadrados a la nube de las n observaciones, habiendo definido así el *hiperplano de regresión ortogonal* (hiperplano de $p - 1$ dimensiones).

Se puede generalizar el análisis mediante la búsqueda de un *subespacio* de regresión ortogonal de $p-q$ dimensiones. Este plano estará caracterizado por ser ortogonal a los q vectores propios de S asociados a los q menores valores propios. Estos vectores propios sucesivos definirán una sucesión de combinaciones lineales de las variables, incorrelacionadas, y de varianza mínima.

INTERPRETACIÓN GEOMÉTRICA DEL ANÁLISIS EN COMPONENTES PRINCIPALES

Se puede realizar una representación gráfica de la nube de puntos X_1, \dots, X_p , columnas de la matriz $X_{(n,p)}$ de casos-variables resultante de medir las variables X_1, \dots, X_p sobre una muestra de tamaño n . Si únicamente se consideran dos variables, una nube de puntos en un plano puede ser asimilada, de forma simple, a una elipse como envolvente de todos ellos. Si se manejaran tres variables, la Figura en relieve resultante sería un elipsoide. Y si fueran más de tres (p), sería necesario un esfuerzo de generalización para imaginar un hiperelipsoide p -dimensional. En general, los ejes principales de estas Figuras o cuerpos geométricos presentarán una inclinación espacial cualquiera respecto a los ejes que representan las variables originales, es decir, no tienen por qué ser paralelos a éstos. Si lo fueran, en un caso hipotético, las proyecciones del elipsoide sobre los planos que definen dos a dos las variables (Figura de la página siguiente) representarían la máxima dispersión de los puntos de la nube (máximas «sombras» del elipsoide) y las variables originales, cada una por sí sola, contendrían la máxima información de la nube en una de las p dimensiones.



Como, además, los ejes son perpendiculares, la dispersión que condensa una de las variables no tiene nada que ver con la que condensan las demás, de modo que cada una recoge la parte de información que no pueden recoger las otras, como ocurre fotografiando un objeto desde tres direcciones perpendiculares.

Pero como la nube, en general, es espacialmente oblicua, las variables originales no recogen solas la información optimizada (máxima «sombra» o dispersión); y habrá otros p ejes (también entre sí perpendiculares y transformados de los ejes iniciales), que serán precisamente los ejes reales del elipsoide, que sí optimizarán la información. Las ecuaciones de estos nuevos ejes serán, como siempre en geometría analítica, combinaciones lineales de las p variables iniciales, que los posicionarán en el espacio. Y se calcularán bajo las condiciones matemáticas de que pasen por el centro de gravedad de la nube, que sean perpendiculares entre sí, que el primero de ellos (el más largo) haga máxima (derivada = 0) la dispersión de la nube sobre él, que el segundo (el siguiente en importancia) haga máxima la dispersión en un plano perpendicular al primero, y así los demás, hasta obtener los p nuevos ejes. Para dos (y hasta para tres) variables iniciales, este cálculo es abordable por métodos habituales de geometría analítica. Pero cuando son más de tres, se llega a tal magnitud de cálculos que es preciso sistematizar la técnica matemática convencional recurriendo a la única herramienta que lo hace posible: el cálculo matricial.

Si los nuevos ejes obtenidos se cortan, como debe ser, en el centro de gravedad de la nube, sus ecuaciones, al no pasar por el origen 0, tendrán un término constante (ordenada en el origen). Pero, como a efectos de información recogida en las proyecciones del elipsoide, no importa más que la dirección de ellos, se puede obligar a que pasen por 0 con lo que todos sus términos independientes desaparecen.

Sus ecuaciones representan nuevas variables ficticias, combinación lineal de las reales, que deberán ser interpretadas, y, si tienen sentido, etiquetadas o nombradas. Se trata de variables sintéticas (**componentes principales**) que, aunque no han sido medidas en los individuos, sí pueden ser calculadas a través de los valores que éstos presenten en todas las variables originales. En cada una de estas p componentes principales habrá algunas variables que contribuyan más a su configuración, o sea, «pesarán» más en su diseño. Este peso viene dado por los valores de los coeficientes de las variables originales en la ecuación de la componente. Y otras variables pesarán menos, revelando una menor influencia de ellas en la componente. Algunas de estas componentes, precisamente las últimas que surjan en el cálculo de la progresivamente menor dispersión perpendicular, pueden proyectar escasa dispersión de la nube desde esas direcciones. (El elipsoide, si fueran tres dimensiones, puede ser tan «aplanado», visto desde la última componente, que considerarlo como una elipse al ser tan estrecho hará perder escasa información). Desechando estas últimas componentes, previa cuantificación, que permite el cálculo, de la poca información perdida, se habrá simplificado la dimensión del problema que pasa de p variables iniciales a k , si bien cada una de las componentes conservadas mantiene en su ecuación a todas las variables originales.

Se puede ver así al elipsoide p -dimensional que contiene toda la información del problema estudiado, desde unas cuantas (k) direcciones principales que presentan k puntos de vista simplificados que el investigador deberá, si puede, interpretar. La adecuada interpretación de estas nuevas variables sintéticas va a depender de que cada una de ellas agrupe con más peso algunas de las variables originales de significado parecido, y con menos peso, las demás. (Por ejemplo, un componente que apareciera con coeficientes altos para la talla y el peso, y bajos para la presión arterial y el colesterol, podría definirse o etiquetarse como una nueva variable llamada tamaño, que no ha sido medida en el estudio).

Pero no siempre será tan destacada la selección útil de variables en las ecuaciones de los componentes principales y, en consecuencia, no resultará fácil la identificación o interpretación de los k «puntos de vista». En tal caso, se debe sacrificar la ubicación óptima de los componentes (ejes ideales del elipsoide), haciéndolos rotar algo (poco, o incluso mucho) para que, perdiendo en la dispersión que cada uno explica, pero no en la dispersión total explicada, ganen en agrupación verosímil de variables y, por tanto, en interpretabilidad práctica. Esta *rotación* última del proceso puede mejorar mucho la utilidad del Análisis de componentes principales en la consecución perseguida de un equilibrio entre la reducción de las dimensiones del problema estudiado y la más fácil interpretabilidad de lo que se conserva.

Esta idea intuitiva geométrica es equivalente a la representación matemática que hemos utilizado para la obtención de las componentes principales Z_1, \dots, Z_p , como combinación lineal de las p variables iniciales X_1, \dots, X_p de la forma siguiente:

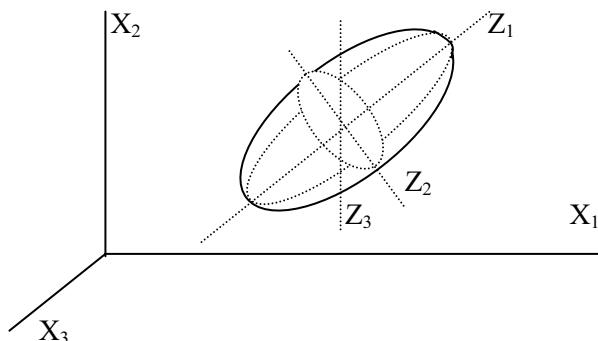
$$\begin{aligned} Z_1 &= u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Z_2 &= u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ &\vdots \\ Z_p &= u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{aligned}$$

donde los u_{ij} representan los *pesos o cargas factoriales* de cada variable en cada componente. Existirán tantos componentes Z_1, \dots, Z_p como número de variables, definidas por p series de coeficientes $u_1 = (u_{11}, \dots, u_{1p})$, ..., $u_p = (u_{p1}, \dots, u_{pp})$. Cada componente explica una parte de la varianza total, considerada ésta como una manera de valorar la información total de la tabla de datos. Si se consigue encontrar pocos componentes (k), capaces de explicar casi toda la varianza total, podrán sustituir a las variables primitivas con mínima pérdida de información. De esta forma se dispondrá de unas variables ficticias que, siendo pocas, contienen a todas las originales. Este es el objetivo del análisis en componentes principales: simplificación o reducción de la tabla inicial, de $n \times p$ a $n \times k$. Naturalmente, si en vez de seleccionar k componentes principales, se tomaran los p posibles, no existiría pérdida alguna de información, pero no se habría conseguido simplificar el problema.

El hiperelipsoide de concentración

Según hemos visto anteriormente, una ***interpretación geométrica clásica de los componentes principales*** puede consistir en la sustitución de la nube de puntos por la elipse que mejor se ajusta (si fueran dos variables), por el elipsoide (si fueran tres) o por «hiper-elipsoides» (si fueran más de tres). Los ejes principales de estos hiperelipsoides corresponderían a las componentes principales, con centro en el centro de gravedad de la nube, ya que recogen la mayor inercia o dispersión de las proyecciones de la nube original de datos sobre ellos. Los autovectores asociados a las componentes definen las direcciones de los ejes principales del hiperelipsoide que encierra la nube de puntos en el espacio.

La primera componente (eje principal Z_1) hace máxima la inercia de la nube de puntos proyectada sobre él. La segunda componente (eje principal Z_2) hace máxima la inercia de la nube proyectada sobre él y no sobre el primero, puesto que son perpendiculares. Todas las demás componentes, tantas como variables, se obtienen de forma correlativa, manteniendo el criterio de perpendicularidad (en espacio multidimensional) entre todas ellas, y sobre cada componente se proyecta la parte de dispersión que no podría proyectarse sobre ningún otro. La nube real de puntos, hasta ahora asimilada a un elipsoide que la envuelve, no adoptará en general esta forma tan regular, por lo que sus ejes principales, ya no definidos geométricamente, deberán ser ubicados por optimización matemática. A continuación se presenta un elipsoide de concentración en el caso de tres dimensiones.



El primer eje principal, que como todos los demás ha de pasar por el centro de gravedad de la nube, será aquél que haga máxima (derivada = 0) la inercia de la nube de puntos proyectada sobre él. Una vez fijado, se elegirá el segundo de entre todas las posiciones que en el espacio puede tomar un eje perpendicular al primero por el centro de gravedad, bajo la condición de que la restante dispersión de la nube sobre él proyectada sea máxima. Y, así, sucesivamente.

La magnitud de los cálculos implicados en este proceso obliga a recurrir al cálculo matricial, herramienta matemática habitual que con la ayuda de técnicas automáticas de proceso de datos facilita la resolución de estos problemas en espacios multidimensionales. El cálculo ha demostrado, además, que la varianza (dispersión relativa) de la nube de puntos explicada por cada componente optimizada es el concepto matricial denominado valor propio (λ) asociado a esa componente. Sabiendo que la *inercia total de la nube* es la suma de las varianzas de las p variables, la proporción de varianza total que recoge cada componente (*porcentaje de inercia explicada por la componente principal h -ésima*) será el porcentaje que representa su valor propio frente a este total, es decir:

$$\frac{\lambda_k}{\lambda_1 + \dots + \lambda_p} \times 100$$

También se puede considerar el *porcentaje de inercia explicada por las k primeras componentes principales (ejes factoriales o factores)* que se define como:

$$\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p} \times 100$$

Ahora bien, como generalmente se parte de la matriz de correlaciones (matriz de covarianzas de las variables estandarizadas), y como la varianza de cada variable estandarizada es 1, la varianza total de la nube será igual a p . En este caso, la proporción de varianza total que recoge cada componente será el porcentaje que representa su valor propio frente a este total, es decir:

$$\frac{\lambda_k}{p} \times 100$$

El porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales) será ahora:

$$\frac{\lambda_1 + \dots + \lambda_k}{p} \times 100$$

Una vez definidos los nuevos ejes perpendiculares o componentes, se pueden calcular las nuevas coordenadas de los puntos de la nube sobre ellos, obteniéndose así una nueva tabla de casos-componentes ($n \times p$), todavía de la mismas dimensiones que la original. Cada coordenada de un caso sobre uno de los ejes se calcula por la función lineal de todas las variables originales. Hemos visto que el cálculo matricial permite la obtención conjunta inmediata de estas coordenadas. Las proyecciones de todos los casos sobre cada nuevo eje tienen, lógicamente, media cero y varianza igual al valor propio relativo a ese eje. El hecho de que las componentes principales estén centradas sobre la nube implica que las proyecciones de los puntos sobre ellas se repartan a ambos lados del origen, y explica que aparezcan valores positivos y negativos. Son, pues, distancias relativas a efectos comparativos, ya que se trata de proyecciones de datos centrados respecto a la media.

Hasta este momento el proceso se limita a definir unos nuevos ejes perpendiculares que sustituyen a los de las variables primitivas y, sobre ellos, una nueva tabla de datos. Sin embargo, este proceso no simplifica la dimensión del problema. Un caso extremo puede ayudar a entender la posibilidad de simplificación: Imagínese que el elipsoide de la Figura anterior fuera muy aplastado, en cuyo caso uno de sus ejes sería muy corto (muy poca dispersión). En el proceso de cálculo descrito, ese componente habría sido el último en ser obtenido.

Su eliminación, por consiguiente, convertiría el elipsoide casi plano en una elipse, prácticamente con el mismo contenido informativo pero con una dimensión menos. Se trata por tanto de seleccionar k de entre estos p componentes, de modo que, la reducción de la dimensión, no suponga una excesiva pérdida de información. El problema está en cuántos componentes retener. La respuesta (no única) va a depender, tal y como ya se ha estudiado anteriormente, de las características del fenómeno estudiado, de la precisión exigida y, sobre todo, de la posibilidad y verosimilitud de interpretación de las componentes principales retenidas, equilibrio no siempre fácil de conseguir, y para el cual el investigador debe esmerar su sentido crítico.

De todas formas, se recomienda el seguimiento de unas directrices basadas en primer lugar en la retención de aquellos componentes cuyo valor propio, calculado a partir de la matriz de correlaciones, sea mayor que 1, lo que significa que explican más varianza que cualquier variable original estandarizada. Así se habrán elegido componentes mejores que variables en capacidad explicativa. En segundo lugar puede adoptarse como directriz la retención de cuantos componentes sean precisos para garantizar conjuntamente un mínimo porcentaje, preestablecido por el investigador, de la dispersión global de la nube. Incluso pueden adoptarse como directrices la retención de los componentes que, individualmente, superen un porcentaje preestablecido o la retención de un número fijo de componentes, independientemente de su capacidad explicativa.

Matriz de cargas factoriales, communalidad y círculos de correlación

La dificultad en la interpretación de los componentes estriba en la necesidad de que tengan sentido y midan algo útil en el contexto del fenómeno estudiado. Por tanto, es indispensable considerar el peso que cada variable original tiene dentro del componente elegido, así como las correlaciones existentes entre variables y factores. Un componente es una función lineal de todas las variables, pero puede estar muy bien correlacionado con algunas de ellas, y menos con otras. Ya hemos visto que el coeficiente de correlación entre una componente y una variable se calcula multiplicando el peso de la variable en esa componente por la raíz cuadrada de su valor propio:

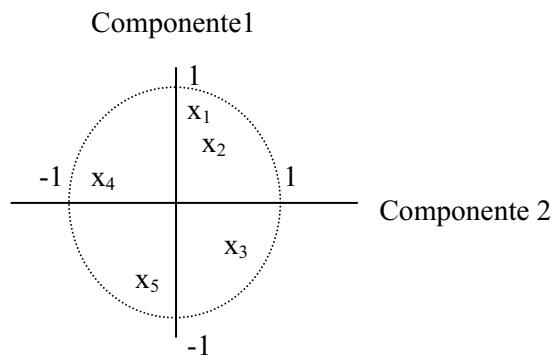
$$r_{jh} = u_{hj} \sqrt{\lambda_h}$$

Se demuestra también que estos coeficientes r representan la parte de varianza de cada variable que explica cada factor. De este modo, cada variable puede ser representada como una función lineal de los k componentes retenidos, donde los pesos o cargas de cada componente o factor (*cargas factoriales*) en la variable coinciden con los coeficientes de correlación. El cálculo matricial permite obtener de forma inmediata la tabla de coeficientes de correlación variables-componentes ($p \times k$), que se denomina **matriz de cargas factoriales**. Las ecuaciones de las variables en función de las componentes (factores), traspuestas las inicialmente planteadas, son de mayor utilidad en la interpretación de los componentes, y se expresan como sigue::

$$\begin{aligned} Z_1 &= r_{11}X_1 + \cdots + r_{1p}X_p & X_1 &= r_{11}Z_1 + \cdots + r_{k1}Z_k \\ Z_2 &= r_{21}X_1 + \cdots + r_{2p}X_p & \Rightarrow & X_2 = r_{12}Z_1 + \cdots + r_{k2}Z_k \\ &\vdots && \vdots \\ Z_k &= r_{k1}X_1 + \cdots + r_{kp}X_p & X_p &= r_{1p}Z_1 + \cdots + r_{kp}Z_k \end{aligned}$$

Por las propiedades del coeficiente de correlación se deduce que la suma en horizontal de los cuadrados de las cargas factoriales de una variable en todos los factores (componentes) retenidos es la parte de dispersión total de la variable explicada por el conjunto de k componentes. Esta suma de cuadrados se denomina **comunalidad**. Por ejemplo, para la primera variable, la communalidad será $r_{11}^2 + \dots + r_{k1}^2 = V(X_1) = h_1^2$. Por consiguiente, la suma de las communalidades de todas las variables representa la parte de inercia global de la nube original explicada por los k factores retenidos, y coincide con la suma de los valores propios de estas componentes. La communalidad proporciona un criterio de calidad de la representación de cada variable, de modo que, variables totalmente representadas tienen de communalidad la unidad. También se demuestra que la suma en vertical de los cuadrados de las cargas factoriales de todas las variables en un componente es su valor propio. Por ejemplo, el valor propio del primer componente será $r_{11}^2 + \dots + r_{1p}^2 = \lambda_1$. Todas estas demostraciones se realizarán de modo formal en el capítulo siguiente.

Es evidente que, al ser las cargas factoriales los coeficientes de correlación entre variables y componentes, su empleo hace comparables los pesos de cada variable en la componente y facilita su interpretación. En este mismo sentido, su representación gráfica puede orientar al investigador en una primera aproximación a la interpretación de los componentes. Como es lógico, esta representación sobre un plano sólo puede contener los factores de dos en dos, por lo que se pueden realizar tantos gráficos como parejas de factores retenidos. Estos gráficos se denominan **círculos de correlación**, y están formados por puntos que representan cada variable por medio de dos coordenadas que miden los coeficientes de correlación de dicha variable con los dos factores o componentes considerados. Todas las variables estarán contenidas dentro de un círculo de radio unidad.



Rotación de las componentes

Es frecuente no encontrar interpretaciones verosímiles a los factores (componentes) obtenidos, ya que se ha organizado el estudio partiendo de un primer componente principal que condensaba la máxima inercia de la nube.

Sin embargo, no tiene por qué coincidir esta máxima inercia del primer factor, que condiciona el cálculo de los restantes, con la óptima interpretación de cada uno de los componentes. Sería deseable, para una más fácil interpretación, que cada componente estuviera muy bien relacionada con pocas variables (coeficientes de correlación r próximos a 1 ó -1) y mal con las demás (r próximos a 0). Esta optimización se obtiene por una adecuada **rotación de los ejes** que definen los componentes principales. Sería tolerable sacrificar la no coincidencia del primero de ellos con el eje principal del elipsoide, si así mejorara la interpretabilidad del conjunto. Rotar un conjunto de componentes no cambia la proporción de inercia total explicada, como tampoco cambia las comunidades de cada variable, que no son sino la proporción de varianza explicada por todos ellos. Sin embargo, los coeficientes, que dependen directamente de la posición de los componentes respecto a las variables originales (cargas factoriales y valores propios), se ven alterados por la rotación.

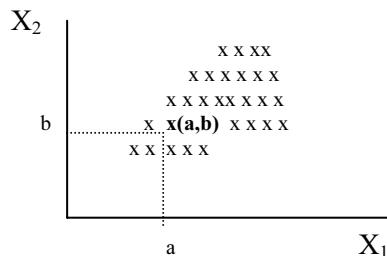
Se puede comprobar que la rotación así descrita equivale a rotar los componentes en el círculo de correlación. Por ejemplo, en el Gráfico anterior puede observarse cómo una rotación que acercara la primera componente a las variables x_1 y x_2 conseguiría una proyección (carga factorial) máxima sobre ella de dichas variables, y mínima en la componente 2 que se mantendría perpendicular al ser arrastrada en el giro. Los nuevos ejes rotados tendrán una correlación con los correspondientes primitivos, que será menor cuanto mayor sea el ángulo de giro.

Existen varios tipos de rotaciones, que serán analizadas en profundidad en el próximo capítulo. Las más utilizadas son la rotación VARIMAX y la QUARTIMAX. La rotación VARIMAX se utiliza para conseguir que cada componente rotado (en vertical, en la matriz de cargas factoriales) presente altas correlaciones sólo con unas cuantas variables, rotación a la que suele aplicarse la llamada *normalización de Kaiser* para evitar que componentes con mayor capacidad explicativa, que no tienen por qué coincidir con la mejor interpretabilidad, pesen más en el cálculo y condicione la rotación. Esta rotación, la más frecuentemente utilizada, es adecuada cuando el número de componentes es reducido. La rotación QUARTIMAX se utiliza para conseguir que cada variable (en horizontal, en la matriz de cargas factoriales) tenga una correlación alta con muy pocos componentes cuando es elevado el número de éstos. Tanto Varimax como Quartimax son **rotaciones ortogonales**, es decir, que se mantiene la condición de perpendicularidad entre cada uno de los ejes rotados. Sin embargo, cuando las componentes, aún rotadas ortogonalmente, no presentan una clara interpretación, cabe todavía la posibilidad de intentar mejorarla a través de **rotaciones oblicuas**, que no respetan la perpendicularidad entre ellos. Piénsese en espacios multidimensionales para comprender la complejidad de los cálculos necesarios. De entre las diversas rotaciones oblicuas desarrolladas, la PROMAX, aplicada normalmente sobre una VARIMAX previa, es la más utilizada dada su relativa simplicidad.

En las soluciones oblicuas varían, lógicamente, no sólo los valores propios sino también las comunidades de las variables y se mantiene, por supuesto, la varianza explicada por el modelo. Además, es importante tener en cuenta que la no perpendicularidad entre los ejes surgida tras una rotación oblicua produce una correlación entre ellos antes inexistente, por lo que la parte de varianza de una variable explicada por una componente no es ya independiente de los demás factores. Deberá valorarse esta relación en la interpretación de los componentes. No se puede decir que una rotación sea mejor que otra, ya que desde un punto de vista estadístico todas son igualmente buenas. La elección entre diferentes rotaciones se basa en criterios no estadísticos, ya que la rotación preferida es aquella más fácilmente interpretable. Si dos rotaciones proponen diferentes interpretaciones no deben ser consideradas discordantes sino como dos enfoques diferentes de un mismo fenómeno que el investigador deberá analizar. La interpretación de una componente es un proceso subjetivo al que la rotación puede restar parte de subjetividad.

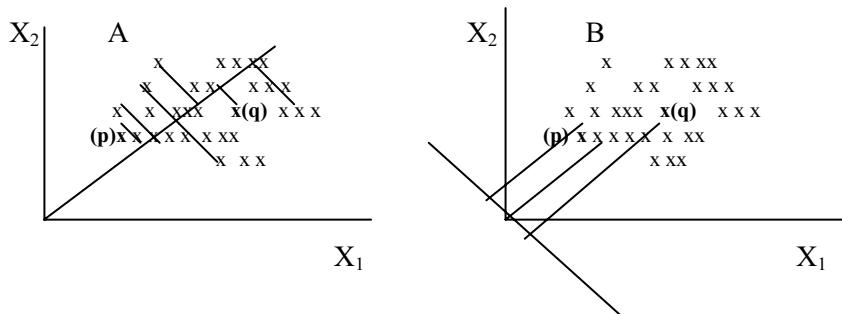
El caso de dos variables

El análisis de las componentes principales constituye un caso de reducción de variables, que visto geométricamente, es un problema de reducción de dimensión. Supóngase, para facilitar su representación gráfica, dos variables aleatorias X_1 y X_2 , por ejemplo edad y presión arterial sistólica. Una serie de observaciones de estas variables se puede representar en una gráfica bidimensional por una nube de puntos, donde cada punto representa una observación. El problema de las componentes principales se puede plantear del siguiente modo: ¿Se puede encontrar una variable, función lineal de ambas, que represente adecuadamente la "información" contenida en las dos? La Figura siguiente presenta el conjunto de observaciones de las variables X_1 y X_2 , donde los valores que toman las variables en cada observación (coordenadas de un punto de la gráfica) son las proyecciones ortogonales sobre los ejes respectivos.



Una nueva variable, función lineal de ambas (por ejemplo, la recta de regresión que ajusta la nube de puntos), será un nuevo eje en el mismo plano y, el valor que esa variable tome para cada observación, será la proyección del punto correspondiente sobre dicho eje. Se tratará, por tanto, de encontrar, si es posible, el eje más adecuado, es decir, aquél sobre el que se obtenga una representación lo más parecida posible a la nube original, esto es, aquél donde las distancias entre los puntos, que indican las diferencias entre las observaciones, mejor se conserven.

En la Figura que se presenta a continuación se observan dos ejes de ajuste para la misma nube de puntos. Comparando las Gráficas *A* y *B* de la Figura, parece más adecuado el eje representado en *A* que el representado en *B*. Los puntos *p* y *q*, que en la representación original están muy alejados entre sí (corresponden a individuos con edades y presiones arteriales muy diferentes), se mantienen más alejados en el eje *A* que en el *B* y, en general, ocurre lo mismo con cualquier par de puntos, aunque siempre podremos encontrar algún par específico para el que la distancia se mantenga mejor en el eje *B*.

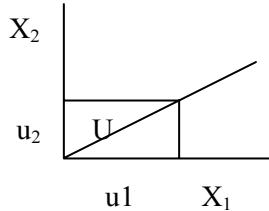


El eje *A* representa mejor la dispersión de los datos originales, y esto establece un criterio para definir el mejor eje como aquél sobre el que la varianza proyectada (medida de la dispersión) sea máxima. La siguiente cuestión a considerar es si este eje representa adecuadamente la dispersión original. Una medida de la dispersión del conjunto de variables originales es la suma de las varianzas de cada una de las variables, habitualmente llamada *inercia*.

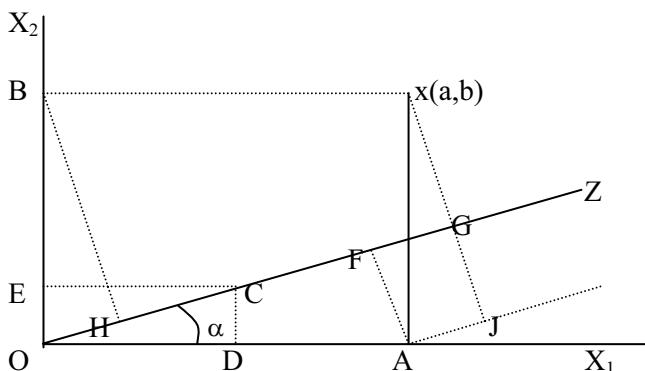
Por consiguiente, una medida de la calidad de la representación que la nueva variable consigue, es la razón entre la varianza proyectada y la inercia inicial (la *inercia total de la nube* es la suma de las varianzas de las p variables X_1, \dots, X_p). En este ejemplo se ha podido encontrar un eje mejor que otros, en términos de máxima varianza, porque las variables están correlacionadas (en la Gráfica, los puntos tienden a estar en la dirección de la recta de regresión). Si las variables no estuvieran correlacionadas, los puntos tendrían una disposición homogénea en el plano y no habría ningún eje sobre el que la varianza proyectada fuese mayor que sobre otros (un eje paralelo a la recta de regresión mantendría toda la varianza). El mejor eje será, por lo tanto, aquél sobre el que se proyecte una mayor cantidad de varianza, y en este sentido podremos concretar el término *Información* para identificar la información contenida en una variable con su varianza.

Una vez hallado el nuevo eje de ajuste, es decir, la variable que mejor resume o *reduce* a las dos iniciales X_1 y X_2 , vamos a calcular el valor que toma dicha variable para cada punto u observación en el plano, que ya sabemos que es la proyección del punto sobre el nuevo eje.

Un eje se define por un vector unitario U que tiene la dirección del eje, y cuyo módulo es la unidad. Si u_1 y u_2 son sus componentes en los ejes X_1 y X_2 , se tiene que $u_1^2 + u_2^2 = 1$, que en notación matricial puede expresarse como $U'U=1$, siendo $U=(u_1,u_2)'$. La Figura siguiente aclara lo expuesto.



Una observación con respecto a las variables X_1 y X_2 , se representa por un punto $x(a,b)$ cuyas coordenadas en el plano son a y b , es decir, $x = (a,b)'$. La coordenada de ese punto respecto al nuevo eje será $z = U'x$ (producto escalar de los vectores U y x) y representa la proyección del punto sobre el nuevo eje. La Figura siguiente aclara estos conceptos.



Las coordenadas del punto x en los ejes originales son, respectivamente, $a = OA$ y $b = OB$. Las coordenadas del vector unitario U ($OC = 1$) en los ejes originales son $u_1 = OD$ y $u_2 = OE$. La coordenada del punto x en el nuevo eje Z es $z = OG = OF+FG$. En el triángulo OFA (rectángulo en F) se tiene que $\cos(\alpha) = OF/OA$, y en el triángulo ODC (rectángulo en D) se tiene que $\cos(\alpha) = OD/OC = OD$, con lo que $OF/OA = OD \Rightarrow OF = OD*OA = u_1a$. Realizando el mismo razonamiento en los triángulos OHB y OEC se encuentra que $OH = OB*OE = u_2b$. Ahora bien, los triángulos OBH y AXJ son iguales (ambos son rectángulos, el ángulo B es igual al ángulo X por estar comprendidos entre paralelas, y el lado OB es igual a AX) con lo que $OH = AJ$. Además, como $FGJA$ es un rectángulo, entonces $AJ = FG$, con lo que $OH = FG$. Por consiguiente $z = OF+FG = u_1a+u_2b$, que en notación matricial puede expresarse como sigue:

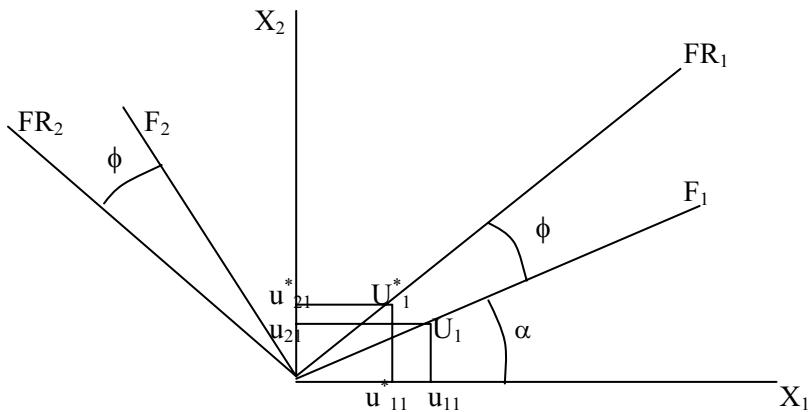
$$z = U'x = (u_1 \ u_2) \begin{pmatrix} a \\ b \end{pmatrix} = u_1 a + u_2 b$$

Por lo tanto, de la interpretación geométrica de las componentes principales se deduce que dada la variable bidimensional $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, se trata de buscar una nueva variable $Z = U'X$, siendo $U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$ con la condición $U'U = 1$, de tal manera que la varianza de Z sea máxima. Pero $Var(Z) = U'\Sigma U$ siendo Σ la matriz de varianzas covarianzas de X , con lo que el problema de componentes principales es encontrar $Z = U'X$ maximizando $U'\Sigma U$ con la condición de que $U'U = 1$. A partir de aquí ya se consideraría todo el aparato algebraico presentado anteriormente para la obtención matemática de las componentes principales.

La principal dificultad del análisis de componentes principales, supuesto un programa de computador con un algoritmo eficiente para diagonalizar matrices, es la interpretación de los factores. Si un factor presentara correlaciones parecidas con muchas variables sería difícil interpretarlo y, generalmente, esto es lo que ocurre. Para resolver esta dificultad se pueden "girar" los factores hasta conseguir que se "parezcan" a alguna variable y así facilitar su interpretación. Esto, en términos no geométricos, significa generar otros factores que, presenten coeficientes de correlación lo más altos posibles con alguna de las variables. Volviendo a la interpretación geométrica de los factores como un sistema de ejes ortogonales, la rotación de los factores es, simplemente, un giro de dichos ejes.

Básicamente hay dos modos de rotación esenciales: *ortogonal* en que los factores se mantienen ortogonales después de la rotación y, por tanto, no garantiza que todos sean fácilmente interpretables y *oblicua* en que cada factor se gira por separado, garantizándose la máxima interpretabilidad de cada uno de ellos pero, a cambio, perdiéndose la independencia entre ellos.

Dentro de la rotación ortogonal, que es la más usada, hay varios métodos dependiendo del criterio con el que se selecciona el ángulo de giro, entre otros: *varimax* que trata de conseguir que cada factor tenga una correlación alta con unas pocas variables y *quartimax* que trata de conseguir que cada variable tenga una correlación alta con unos pocos factores. Para facilitar la interpretación geométrica de la rotación de los factores, realizaremos la representación gráfica para el caso de dos variables X_1 y X_2 y dos factores F_1 y F_2 que se presenta en la Figura siguiente:



Una rotación ortogonal está definida por el ángulo ϕ que gira cada factor (el mismo para todos). Habrá que encontrar una transformación que convierta el vector U_1 , de componentes u_{11} y u_{21} y que define el eje del factor F_1 , en el vector U^*_1 , de componentes u^*_{11} y u^*_{21} y que define el factor rotado FR_1 . Esta transformación será, en una rotación ortogonal, la misma para todos los factores. Teniendo en cuenta que tanto U_1 , como U^*_1 , son unitarios tenemos lo siguiente:

$$\begin{aligned} u_{11} &= \cos(\alpha) & u^*_{11} &= \cos(\alpha + \phi) = \cos(\alpha)\cos(\phi) - \sin(\alpha)\sin(\phi) \\ u_{21} &= \sin(\alpha) & u^*_{21} &= \sin(\alpha + \phi) = \sin(\alpha)\cos(\phi) + \cos(\alpha)\sin(\phi) \end{aligned}$$

con lo que se tiene:

$$\begin{aligned} u^*_{11} &= u_{11}\cos(\phi) - u_{21}\sin(\phi) \\ u^*_{21} &= u_{11}\sin(\phi) + u_{21}\cos(\phi) \end{aligned}$$

que en formato matricial se expresa como sigue:

$$\begin{pmatrix} u^*_{11} \\ u^*_{21} \end{pmatrix} = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix}$$

luego la transformación que convierte el vector U_1 , de componentes u_{11} y u_{21} y que define el eje del factor F_1 , en el vector U^*_1 , de componentes u^*_{11} y u^*_{21} y que define el factor rotado FR_1 , puede expresarse como $U^*_1 = R U_1$, siendo R la matriz de la rotación, que viene dada por:

$$R = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}$$

Cualquier vector se transforma en su rotado multiplicando sus coordenadas por la matriz R . Es de destacar que los nuevos vectores transformados de los vectores propios de la matriz de covarianzas en la rotación ya no son vectores propios de la matriz de covarianzas, y los nuevos valores propios asociados a los vectores propios transformados por la rotación no coinciden con la varianza de los nuevos factores. Sin embargo, las rotaciones ortogonales conservan las comunidades de cada variable, así como la suma de las varianzas de los factores.

La matriz U cuyas columnas son los dos autovectores se transformará en $U^*=RU$, de donde se deduce (multiplicando a la derecha por la matriz ortogonal U') que $U^*U'=R$. En el caso general en que el número de variables p es mayor de dos, la rotación está definida por una matriz similar R cuadrada de orden p , en la que aparecen los cosenos de las proyecciones del ángulo ϕ sobre los distintos planos. Los nuevos factores rotados se calculan mediante:

$$Z^* = (U^*)'X = (RU)'X = U'R'X$$

Las varianzas de los nuevos factores rotados serán:

$$Var(Z_i^*) = (U_i^*)'\Sigma U_i^* = (RU_i)'\Sigma RU_i = U_i'R'\Sigma RU_i$$

y la matriz de varianzas covarianzas con los nuevos factores es:

$$U^* = \Sigma U^* = \Sigma RU$$

El problema esencial en la rotación de los factores es encontrar la matriz R más adecuada, y es en esta tarea donde difieren los distintos métodos de rotación ortogonal. Por ejemplo, el método *varimax* elige la matriz R de modo que sea máxima la suma:

$$S = \sum_{i=1}^k S_i \text{ con } S_i = -\frac{p \sum_{j=1}^p (r_{ji}^2)^2 - \left(\sum_{j=1}^p r_{ji}^2 \right)^2}{p^2} \text{ para } i = 1, \dots, k$$

con el objeto de simplificar cada factor para que sus cargas factoriales (coeficientes de correlación entre los factores y las variables) sean altas sólo con algunas variables y pequeñas con el resto. En este método se maximiza la suma de las cargas factoriales cuadráticas de los factores (o *simplicidades* S_i) que, tal y como se han definido, serán grandes cuando haya cargas factoriales extremas y pequeñas para cargas factoriales con valores próximos.

Una variación de este método es la *normalización de Kaiser*, que es igual al *varimax*, pero dividiendo cada carga factorial cuadrática r_{ij}^2 por la communalidad h_j^2 de la variable X_j . De esta forma se evita que los factores con sumas de cargas factoriales más altas tengan más peso.

Por lo tanto, en el método de *normalización de Kaiser* se elige la matriz R de forma que sea máxima la suma:

$$S = \sum_{i=1}^k S_i \text{ con } S_i = \frac{p \sum_{j=1}^p \left(\frac{r_{ji}^2}{h_j^2} \right)^2 - \left(\sum_{j=1}^p \frac{r_{ji}^2}{h_j^2} \right)^2}{p^2} \text{ para } i = 1, \dots, k$$

Otro método de rotación ortogonal es el método quartimax, que pretende que cada variable tenga una correlación alta con muy pocos factores. Para ello S maximiza la varianza de las cargas factoriales para cada variable, en lugar de para cada factor, con la condición de que se mantengan las comunidades de cada variable. Se demuestra que esto es equivalente a hacer máxima la suma:

$$Q = \sum_{i=1}^k \sum_{j=1}^p r_{ij}^4$$

PROPIEDADES MUESTRALES DE LAS COMPONENTES PRINCIPALES

Supongamos que partimos de N observaciones independientes obtenidas como muestra aleatoria de una normal n-dimensional (μ, Σ) expresadas como sigue:

$$\begin{matrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} & X_{2N} & & X_{nN} \end{matrix}$$

Si designamos la estimación de Σ (con valores propios $\lambda_1, \lambda_2, \dots, \lambda_n$) por S , los vectores propios estimados a partir de S por b_1, b_2, \dots, b_n y los valores propios estimados a partir de S por l_1, l_2, \dots, l_n , se demuestra que cuando N es grande se cumple lo siguiente:

- l_i se distribuye independientemente de los elementos que componen b_i .
- $\sqrt{n}(l_i - \lambda_i)$ se distribuye normalmente con media cero y varianza $2\lambda_i^2$ e independientemente del resto de los autovalores.
- $\sqrt{n}(b_i - u_i)$ se distribuye de acuerdo con una normal multivariante con vector

media cero y matriz de varianzas covarianzas $\lambda_i \sum_{\substack{h=1 \\ h \neq i}}^n \frac{\lambda_h}{(\lambda_h - \lambda_i)^2} u_h u'_h$.

- La covarianza del elemento h-esimo de u_i y del S-ésimo elemento de u_j es:

$$\frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} u_{ih} u_{jS} \quad i \neq j$$

A partir de los resultados anteriores pueden establecerse contrastes de hipótesis y construirse intervalos de confianza. Tenemos los siguientes resultados:

- $\frac{\sqrt{n}(l_i - \lambda_i)}{\sqrt{2}l_i} \rightarrow N(0,1) \Rightarrow l_i \pm z_{\alpha/2} \sqrt{2/n} l_i$ es intervalo de confianza de nivel α de l_i
- La región crítica del contraste $H_0: \lambda_i = \lambda_i^0$ contra $H_1: \lambda_i \neq \lambda_i^0$ es $\left| \frac{\sqrt{n}(l_i - \lambda_i^0)}{\sqrt{2}\lambda_i^0} \right| > z_{\alpha/2}$
- Como $\sum_{i=1}^k \lambda_i$ es la varianza retenida por los k factores, interesa contrastar la hipótesis de que si los factores que no se han retenido realmente explican menos varianza que una cantidad prefijada τ . Se contrastará entonces $H_0: \sum_{i=k+1}^p \lambda_i = \tau$ contra $H_1: \sum_{i=k+1}^p \lambda_i < \tau$, para lo que se usa el estadístico $\sum_{i=k+1}^p \sqrt{n}(l_i - \lambda_i)$ de media 0 y varianza $2 \sum_{i=k+1}^p \lambda_i^2$, siendo la región crítica $\sum_{i=k+1}^p l_i < \tau - z_{\alpha} \sqrt{\frac{2}{n} \sum_{i=k+1}^p l_i^2}$, y siendo un intervalo de confianza para $\sum_{i=k+1}^p \lambda_i$ al nivel α el definido como $\sum_{i=k+1}^p l_i \pm z_{\alpha/2} \sqrt{\frac{2}{n} \sum_{i=k+1}^p l_i^2}$
- Puede contrastarse la hipótesis nula de que la multiplicidad del valor propio λ_k es r , o lo que es lo mismo, que r autovalores intermedios son iguales ($H_0: \lambda_{q+1} = \lambda_{q+2} = \dots = \lambda_{q+r}$) mediante el estadístico $T = -N \sum_{k=q+1}^{q+r} \ln(l_k) + Nr \ln\left(\frac{1}{r} \sum_{k=q+1}^{q+r} l_k\right)$ que se distribuye según una chi-cuadrado con $N \ln\left[\left(\frac{1}{r} \sum_{k=q+1}^{q+r} l_k\right) / \prod_{k=q+1}^{q+r} l_k\right]$ grados de libertad.

CAPÍTULO 5

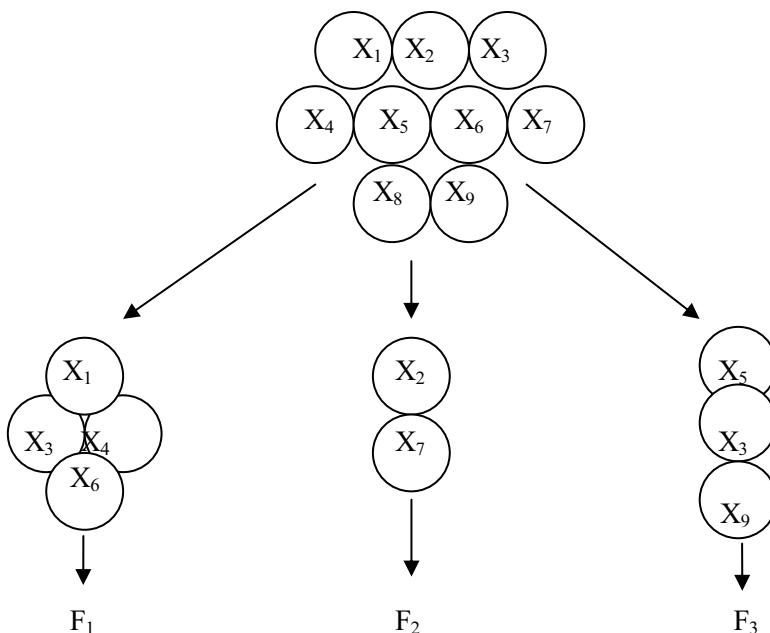
ANÁLISIS FACTORIAL

OBJETIVO DEL ANÁLISIS FACTORIAL

El análisis factorial tiene como objeto simplificar las múltiples y complejas relaciones que puedan existir entre un conjunto de variables observadas $X_1 X_2 \dots X_p$. Para ello trata de encontrar dimensiones comunes o *factores* que ligan a las aparentemente no relacionadas variables. Concretamente, se trata de encontrar un conjunto de $k < p$ *factores no directamente observables* $F_1, F_2 \dots F_k$ que expliquen suficientemente a las variables observadas perdiendo el mínimo de información, de modo que sean fácilmente interpretables (*Principio de interpretabilidad*) y que sean los menos posibles, es decir, k pequeño (*Principio de parsimonia*). Además, los factores han de extraerse de forma que resulten independientes entre sí, es decir, que sean ortogonales. En consecuencia, el análisis factorial es una técnica de reducción de datos que examina la interdependencia de variables y proporciona conocimiento de la estructura subyacente de los datos.

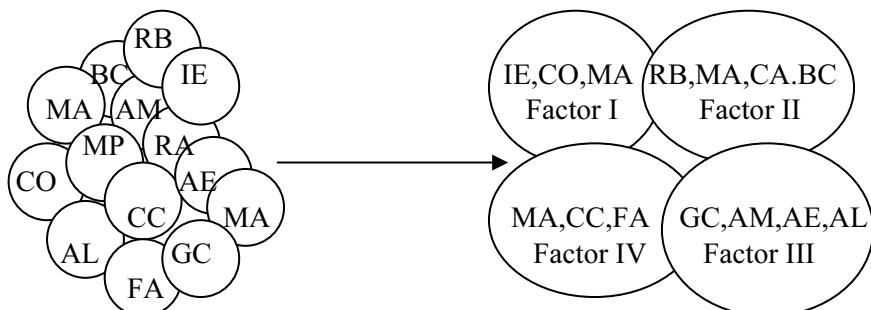
El aspecto más característico del análisis factorial lo constituye su capacidad de reducción de datos. Las relaciones entre las variables observadas $X_1 X_2 \dots X_p$ vienen dadas por su matriz de correlaciones, de modo que, en el análisis factorial se puede partir de una serie de coeficientes de correlación para el conjunto de variables observadas y, a continuación, estudiar si subyace algún patrón de relaciones tal que los datos puedan ser reordenados a un conjunto menor de factores que podemos considerar como variables que recogen y resumen las interrelaciones observadas en los datos.

Como ejemplo ilustrativo, supongamos que tenemos nueve variables $X_1, X_2, \dots X_9$ que se intentan resumir por tres factores no observables F_1, F_2 y F_3 . Analizando las relaciones entre las variables se observa que las variables X_1, X_3, X_4 y X_6 están fuertemente correlacionadas con otra F_1 que, por lo tanto, constituirá el primer factor. De forma similar las variables X_2 y X_7 se agrupan en el segundo factor F_2 y las variables X_5, X_8 y X_9 se agrupan en el tercer factor F_3 . De forma gráfica podríamos expresar este hecho como sigue:



Como aplicación práctica supongamos que se quiere analizar la importancia que los consumidores dan a 14 variables que se consideran relevantes para la compra de un automóvil. Estas variables son: reparaciones baratas (RB), amplia gama de colores (GC), interior espacioso (IE), bajo consumo de gasolina (BC), manejabilidad (MA), aspecto moderno (AM), valor de recompra alto (RA), confortable (CO), motor potente (MP), aspecto elegante (AE), cómodo de conducir (CC), atractivo de línea (AL), maletero amplio (MA) y fácil de aparcar (FA).

Se observa que las 14 variables pueden caracterizarse por cuatro dimensiones subyacentes relacionadas respectivamente con el confort (factor I), con el coste-eficiencia (factor II), con la elegancia (factor III) y con el manejo fácil (factor IV) y no observables directamente. Por lo tanto, en vez de considerar las 14 variables, simplificaremos las cosas, de forma que sólo cuatro factores deban considerarse para caracterizar la estructura subyacente de los datos. De forma gráfica podríamos expresar este hecho como se indica a continuación:



El análisis de componentes principales y el análisis factorial tienen en común que son técnicas para examinar la interdependencia de variables. Difieren en su objetivo, sus características y su grado de formalización, según se verá a continuación.

En el Análisis de componentes principales se obtenían unas variables sintéticas, combinaciones de las originales, cuyo cálculo se basaba únicamente en aspectos matemáticos, independientes de su interpretabilidad práctica que más tarde sería analizada. Si no fueran interpretables, se habrían conseguido unas variables ficticias inútiles para la investigación, aunque matemáticamente siempre calculables. Si lo fueran, se habrían encontrado nuevas variables no medidas pero biológicamente útiles que han aflorado, sin saber lo que se buscaba, a partir de meras relaciones matemáticas entre las variables originales. En el Análisis factorial se presupone la existencia de ciertas variables no medidas y de interés biológico que, latentes en la tabla de datos, permanecen a la espera de ser halladas. Esta presunción de existencia de variables subyacentes es la condición clave del Análisis factorial. Se trata de un método estadístico multivariante distinto del Análisis de componentes principales aunque con soporte matemático parecido, que trata de encontrar variables sintéticas latentes e inobservables, cuya existencia se sospecha. Desde este punto de vista, también acaba siendo un método de simplificación o reducción de la complejidad de la tabla de casos variables con datos cuantitativos, aunque no es éste su objetivo último.

Recuérdese que en Componentes principales, por haber tantos ejes como variables antes de la retención de los mejores, la varianza de las variables originales quedaba totalmente explicada por ellos. Mientras que el objetivo del análisis de componentes principales es explicar la mayor parte de la variabilidad total de un conjunto de variables con el menor número de componentes posible, en el análisis factorial, los factores son seleccionados para explicar las interrelaciones entre variables. En componentes principales se determinan los pesos o ponderaciones que tienen cada una de las variables en cada componente; es decir, las componentes principales se explican en función de las variables observables. Sin embargo, en el análisis factorial las variables originales juegan el papel de variables dependientes que se explican por factores comunes y únicos, que no son observables.

Por otra parte, el análisis de componentes principales es una técnica estadística de reducción de datos que puede situarse en el dominio de la estadística descriptiva, mientras que el análisis factorial implica la elaboración de un modelo que requiere la formulación de hipótesis estadísticas y la aplicación de métodos de inferencia estadística. Como veremos posteriormente, el hecho de que las componentes principales se utilicen como uno de los procedimientos para la extracción de factores, ha podido hacer pensar a algunos erróneamente que son métodos completamente equivalentes. Por otra parte, en algunos programas de computador, por ejemplo en SPSS, ambas técnicas están dentro del mismo procedimiento general.

EL MODELO FACTORIAL

Consideramos las variables observables X_1, X_2, \dots, X_p como variables tipificadas (con media cero y varianza unidad) y vamos a formalizar la relación entre variables observables y factores definiendo el **modelo factorial** de la siguiente forma:

$$\begin{aligned}X_1 &= l_{11}F_1 + l_{12}F_2 + \dots + l_{1k}F_k + e_1 \\X_2 &= l_{21}F_1 + l_{22}F_2 + \dots + l_{2k}F_k + e_2 \\&\vdots \\X_p &= l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pk}F_k + e_k\end{aligned}$$

En este modelo, F_1, F_2, \dots, F_k son los **factores comunes**; e_1, e_2, \dots, e_p son los **factores únicos** o **factores específicos** y l_{jh} es el **peso** del factor h en la variable j , denominado también **carga factorial** o **saturación** de la variable j en el factor h . Según la formulación del modelo, cada una de las p variables observables es una combinación lineal de k **factores comunes** a todas las variables ($k < p$) y de un **factor único** para cada variable. Así pues, todas las variables originales están influenciadas por todos los factores comunes, mientras que para cada variable existe un factor único que es específico para esa variable. Tanto los factores comunes como los específicos son variables no observables. El modelo factorial en forma matricial se expresa como sigue:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1k} \\ l_{21} & l_{22} & \dots & l_{2k} \\ \vdots & & & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pk} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}$$

o lo que es lo mismo:

$$X = LF + e$$

Hipótesis en el modelo factorial

Para poder aplicar la teoría de la inferencia estadística en el modelo factorial es necesario formular hipótesis estadísticas sobre los factores comunes y sobre los factores únicos. Consideraremos los factores comunes F_1, F_2, \dots, F_k como variables tipificadas de media cero y varianza unitaria, y que además no están correlacionadas entre sí. Según esta condición la matriz de covarianzas de los factores comunes es la matriz identidad ($E[FF'] = I$) y la esperanza del vector de factores comunes es el vector cero ($E[F] = 0$).

Por otra parte, se supone que la matriz de covarianzas de los factores específicos (únicos) es una matriz diagonal, lo que implica que las varianzas de los factores únicos pueden ser distintas y que dichos factores únicos están incorrelacionados entre sí, es decir: $E[ee'] = \Omega$ con Ω matriz diagonal. Por otro lado, la esperanza del vector de factores comunes se supone que es el vector cero ($E[e] = 0$).

Por último, se tendrá en cuenta que para poder realizar inferencias que permitan distinguir, para cada variable, entre los factores comunes y el factor único, es necesario suponer que los factores comunes están incorrelacionados con el factor único, es decir, que la matriz de covarianzas entre los factores comunes y los factores únicos es la matriz cero ($E[Fe'] = 0$).

Resumiendo las hipótesis previamente citadas tenemos:

$$\text{Modelo} \rightarrow X = LF + e \quad \text{Hipótesis} \rightarrow E[FF'] = I, E[F] = 0, E[ee'] = \Omega, E[e] = 0, E[Fe'] = 0$$

Comunalidades y especificidades

Dado que las variables X son variables tipificadas, su matriz de covarianzas es igual a la matriz de correlación poblacional R_p , matriz que puede descomponerse de la forma siguiente:

$$R_p = E(XX') = E(LF+e)(LF+e)' = LE(FF')L' + E(ee') + LE(fe') + E(fe)L' = LIL' + \Omega + L0 + 0L' = LL' + \Omega$$

La relación anterior puede expresarse en forma matricial como sigue:

$$\begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & & & \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1k} \\ l_{21} & l_{22} & \cdots & l_{2k} \\ \vdots & & & \\ l_{p1} & l_{p2} & \cdots & l_{pk} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{p1} \\ l_{12} & l_{22} & \cdots & l_{p2} \\ \vdots & & & \\ l_{1k} & l_{2k} & \cdots & l_{pk} \end{bmatrix} + \begin{bmatrix} \varpi_1^2 & 0 & \cdots & 0 \\ 0 & \varpi_2^2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \varpi_p^2 \end{bmatrix}$$

En esta descomposición LL' es la parte correspondiente a los factores comunes y Ω es la matriz de covarianzas de los factores únicos. Además, en la descomposición se observa que la varianza de la variable tipificada X_j se puede expresar como:

$$V_j = 1 = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jp}^2 + \varpi_j^2$$

y si denominamos:

$$h_j^2 = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jp}^2$$

tenemos la descomposición de la varianza poblacional de la variable X_j como:

$$V_j = 1 = h_j^2 + \varpi_j^2 \quad j = 1 \dots p$$

Se observa que h_j^2 es la parte de la varianza de la variable X_j debida a los factores comunes, y se denomina **comunalidad**.

También se observa que ϖ_j^2 es la parte de la varianza de la variable X_j debida a los factores únicos (o específicos), y se denomina **especificidad**.

De la relación matricial anterior también se deduce que la correlación entre cada par de variables originales X_h y X_j viene dada en función de los coeficientes de los factores comunes como sigue:

$$\rho_{hj} = l_{h1}l_{j1} + l_{h2}l_{j2} + \dots + l_{hp}l_{jp} = \sum_{s=1}^p l_{hs}l_{js}$$

MÉTODO DE TURSTONE PARA OBTENER LOS FACTORES

El problema fundamental en el análisis factorial es la estimación de los coeficientes l_{jh} de la **matriz factorial** L en el modelo $X = LF + e$. A las estimaciones de estos coeficientes se les denomina **cargas factoriales estimadas**, aunque en la práctica suele omitirse el calificativo estimadas. Las cargas factoriales estimadas nos indican los pesos de los distintos factores en la estimación de la communalidad de cada variable. Una vez estimado h_j^2 (comunalidad) a partir de las estimaciones de los l_{jh} (cargas factoriales) aplicando que $h_j^2 = l_{j1}^2 + l_{j2}^2 + \dots + l_{jp}^2$, se realiza la estimación de ϖ_j^2 (especificidad) sencillamente por diferencia aplicando que $\varpi_j^2 = 1 - h_j^2$.

Inicialmente, las matrices de la relación $R_p = LL' + \Omega$, a partir de la cual han de calcularse los coeficientes l_{jh} de la matriz factorial L en el modelo $X = LF + e$, están integradas por parámetros poblacionales que son desconocidos. Será necesario entonces utilizar estimaciones lógicas de estas matrices, cuyos elementos sean conocidos a partir de los datos muestrales. Como es natural, estimaremos la matriz de correlación poblacional R_p por la matriz de correlación muestral R, definida como:

$$R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

con lo que la relación $R_p = LL' + \Omega$ pasará a tomar la forma $R = \hat{L}\hat{L}' + \hat{\Omega}$, siendo necesario ahora la obtención de las matrices estimadas \hat{L} y $\hat{\Omega}$ a partir del conocimiento de la matriz de correlación muestral R. La obtención de estas matrices estimadas no es trivial, ya que surgen problemas de ***no unicidad de las soluciones*** y de ***grados de libertad*** en la resolución del sistema $R = \hat{L}\hat{L}' + \hat{\Omega}$.

Las soluciones obtenidas para \hat{L} en el sistema $R = \hat{L}\hat{L}' + \hat{\Omega}$ no tienen por qué ser únicas, ya que si \hat{L} es una solución, también será solución cualquier transformación ortogonal suya $B = \hat{L}H$ con $HH' = 1$. Esto es así porque $BB' = \hat{L}HH'\hat{L}' = \hat{L}\hat{L}'$.

Por otra parte, el sistema $R = \hat{L}\hat{L}' + \hat{\Omega}$ tiene p^2 ecuaciones, que es el número de elementos de R. Pero la matriz R es simétrica y, consecuentemente, está integrada por $p(p+1)/2$ elementos distintos, que será el número real de ecuaciones distintas de las que disponemos. Sin embargo, el número de parámetros a estimar viene dado por los $p \times k$ elementos de la matriz \hat{L} y los p elementos de la matriz $\hat{\Omega}$, esto es, tenemos $p \times k + p = p(k+1)$ parámetros a estimar. En consecuencia, para que el sistema tenga solución posible, es decir, para que se pueda llevar a cabo la estimación, se requiere que el número de ecuaciones sea mayor o igual que el número de parámetros a estimar o incógnitas del sistema. Ha de cumplirse entonces que:

$$p(p+1)/2 \geq p(k+1)$$

Ya que la correlación poblacional o teórica entre cada par de variables originales X_h y X_j viene dada en función de los coeficientes de los factores comunes por la expresión:

$$\rho_{hj} = l_{h1}l_{j1} + l_{h2}l_{j2} + \cdots + l_{hp}l_{jp} = \sum_{s=1}^p l_{hs}l_{js}$$

existirá la correspondiente expresión muestral, que viene dada por:

$$r_{hj} = \hat{l}_{h1}\hat{l}_{j1} + \hat{l}_{h2}\hat{l}_{j2} + \cdots + \hat{l}_{hp}\hat{l}_{jp} = \sum_{s=1}^p \hat{l}_{hs}\hat{l}_{js}$$

La matriz de elementos r_{hj} suele llamarse ***matriz de correlación reproducida***.

Como las variables están tipificadas, la carga factorial \hat{l}_{jf} es el coeficiente correlación muestral entre la variable X_j y el factor F_f . Cuando las variables no están tipificadas la correlación entre la variable X_j y el factor F_f es $\frac{\hat{l}_{jf}}{\sigma(X_j)}$.

MÉTODO DEL FACTOR PRINCIPAL PARA OBTENER LOS FACTORES

Partimos de nuestro modelo factorial:

$$\begin{aligned} X_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1k}F_k + e_1 \\ X_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2k}F_k + e_2 \\ &\vdots \\ X_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pk}F_k + e_k \end{aligned}$$

Considerando las variables X_j reducidas, la varianza total de las p variables X_j será p . De ese total, la varianza explicada por los factores comunes es la suma de las communalidades, y la explicada exclusivamente por el factor F_j es:

$$V_j = l_{1j}^2 + l_{2j}^2 + \cdots + l_{pj}^2$$

Adicionalmente sabemos que:

$$\rho_{hj} = l_{h1}l_{j1} + l_{h2}l_{j2} + \cdots + l_{hp}l_{jp} = \sum_{s=1}^p l_{hs}l_{js} \quad h,j = 1 \dots p$$

pudiendo estimarse el coeficiente de correlación poblacional ρ_{hj} por el coeficiente de correlación muestral r_{hj} .

El método del factor principal obtiene el primer factor maximizando la varianza explicada por él, que es $V_1 = l_{11}^2 + l_{21}^2 + \cdots + l_{p1}^2$, sujeta a las restricciones:

$$r_{hj} = \sum_{s=1}^p l_{hs}l_{js} \quad h,j = 1 \dots p$$

Nos encontramos ante un problema de optimización con restricciones, que se resuelve a partir del método de los multiplicadores de Lagrange considerando la función:

$$G_1 = V_1 + \sum_{h,j=1}^p v_{hj} (r_{hj} - \sum_{s=1}^k l_{hs}l_{js}) \quad v_{hj} = \text{multiplicadores de Lagrange}$$

Derivando la función lagrangiana respecto de las incógnitas (l_{hs}) e igualando a cero, tenemos la expresión fundamental:

$$\frac{\partial G_1}{\partial l_{hs}} = \delta_{1s} l_{h1} - \sum_{j=1}^p v_{hj} l_{js} = 0 \quad s = 1 \dots p \quad \delta_{1s} = \begin{cases} 1 & \text{si } s = 1 \\ 0 & \text{si } s \neq 1 \end{cases}$$

Para $s=1$, en esta expresión fundamental se tiene que $l_{h1} = \sum_{j=1}^p v_{hj} l_{j1}$.

Por otra parte, si en la expresión fundamental multiplicamos a ambos lados por l_{h1} y sumamos respecto a h , tenemos:

$$\delta_{1s} \sum_{h=1}^p l_{h1}^2 - \sum_{j=1}^p \sum_{h=1}^p v_{hj} l_{h1} l_{js} = 0 \quad s = 1 \dots p \quad \delta_{1s} = \begin{cases} 1 & \text{si } s = 1 \\ 0 & \text{si } s \neq 1 \end{cases}$$

Si en esta última expresión hacemos $\sum_{h=1}^p l_{h1}^2 = \lambda_1$ y tenemos en cuenta que:

$$l_{h1} = \sum_{j=1}^p v_{hj} l_{j1} \Rightarrow l_{j1} = \sum_{h=1}^p v_{hj} l_{h1} \quad (v_{hj} = v_{jh}) \text{, ya podemos escribir:}$$

$$\delta_{1s} \lambda_1 - \sum_{j=1}^p l_{j1} l_{js} = 0 \quad s = 1 \dots p \quad \delta_{1s} = \begin{cases} 1 & \text{si } s = 1 \\ 0 & \text{si } s \neq 1 \end{cases}$$

Multiplicando la expresión anterior por l_{hs} y sumando en s , se tiene:

$$l_{h1} \lambda_1 - \sum_{j=1}^p l_{j1} \left(\underbrace{\sum_{s=1}^p l_{hs} l_{js}}_{r_{hj}} \right) = 0 \Rightarrow \sum_{j=1}^p l_{j1} r_{hj} - l_{h1} \lambda_1 = 0 \quad h=1 \dots p$$

Esto es:

$$(h_1^2 - \lambda_1) l_{11} + r_{12} l_{21} + \dots + r_{1p} l_{p1} = 0$$

$$r_{21} l_{11} + (h_2^2 - \lambda_1) l_{21} + \dots + r_{2p} l_{p1} = 0$$

\vdots

$$r_{p1} l_{11} + r_{p2} l_{21} + \dots + (h_n^2 - \lambda_1) a_{n1} = 0$$

Por lo tanto, λ_1 es el mayor valor propio de la matriz de correlaciones LL' y $(l_{11}, l_{21}, \dots, l_{p1})'$ es su vector propio asociado, de módulo λ_1 . Por lo tanto, se tiene que:

$$l_{i1} = \alpha_{i1} \sqrt{\lambda_1} \quad i = 1 \dots p$$

siendo $(\alpha_{11}, \alpha_{21}, \dots, \alpha_{p1})$ un vector propio de módulo unidad y λ_1 el mayor valor propio de la matriz de correlaciones LL' .

Una vez obtenidos los pesos (cargas factoriales o saturaciones) del primer factor, que es el que más contribuye a la varianza de las variables, eliminamos su influencia considerando el nuevo modelo factorial:

$$\begin{aligned} X_1' &= X_1 - l_{11}F_1 = l_{12}F_2 + \dots + l_{1k}F_k + e_1 \\ X_2' &= X_2 - l_{21}F_1 = l_{22}F_2 + \dots + l_{2k}F_k + e_2 \\ &\vdots \\ X_p' &= X_p - l_{p1}F_1 = l_{p2}F_2 + \dots + l_{pk}F_k + e_p \end{aligned}$$

y obtenemos el segundo factor maximizando la varianza explicada por él en este segundo modelo, que es $V_2 = l_{12}^2 + l_{22}^2 + \dots + l_{p2}^2$, sujeta a las restricciones anteriores.

Operando como antes se demuestra que:

$$l_{i2} = \alpha_{i2} \sqrt{\lambda_2} \quad i = 1 \dots p$$

siendo $(\alpha_{12}, \alpha_{22}, \dots, \alpha_{p2})$ un vector propio de módulo unidad y λ_2 el segundo mayor valor propio de la matriz de correlaciones LL' . Ya hemos obtenido los pesos del segundo factor.

Se repite el proceso hasta obtener los pesos de todos los factores, es decir la matriz factorial, al menos hasta que la varianza total explicada por los factores comunes sea igual o próxima a la suma de las communalidades.

El número de factores obtenidos coincide con el de valores propios no nulos de LL' , que son todos positivos ya que LL' es simétrica semidefinida positiva. Hay que tener en cuenta que en la práctica sólo se dispone de correlaciones muestrales, lo que introduce un cierto error de muestreo en el cálculo de los valores propios, error que intenta salvarse fijando una constante positiva c y calculando los valores propios mayores que c , cuyo número indicará el de factores comunes en el modelo factorial. Suele tomarse por lo menos $c=1$ para que la variabilidad explicada por cada factor común supere a la varianza de una variable (que es la unidad).

El método del factor principal puede explicarse por la diagonalización de la matriz LL' , que tomará la forma:

$$LL' = TD_\lambda T'$$

siendo T la matriz cuyas k columnas son los vectores propios de módulo unidad de LL' y siendo $D_\lambda = \text{diag}(\lambda_1 \dots \lambda_k)$. La matriz factorial será entonces:

$$L = TD_\lambda^{1/2}$$

MÉTODO ALPHA PARA OBTENER LOS FACTORES

Este método determina la matriz factorial especificando un número k de factores comunes y formando la matriz T de dimensión $p \times k$ con los autovectores unitarios correspondientes a los k primeros vectores propios de la matriz $H^{-1} LL' H^{-1}$, siendo H^2 la matriz de communalidades (matriz con las communalidades en la diagonal principal).

Si $D_\lambda = \text{diag}(\lambda_1 \dots \lambda_k)$ entonces, al diagonalizar la matriz $H^{-1} LL' H^{-1}$, se tiene:

$$H^{-1} LL' H^{-1} = T D_\lambda T' \Rightarrow LL' = HT D_\lambda T' H^{-1} \quad (H' = H)$$

Con lo que la matriz factorial será $L = H TD_\lambda^{1/2}$

MÉTODO DEL CENTROIDE PARA OBTENER LOS FACTORES

En el método del centroide se elige el primer factor de modo que pase por el centro de gravedad (**centroide**) de las variables sin unicidades $X_i' = X_i - e_i$. Tenemos entonces el modelo factorial:

$$\begin{aligned} X_1' &= X_1 - e_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1k}F_k \\ X_2' &= X_2 - e_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2k}F_k \\ &\vdots \\ X_p' &= X_p - e_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pk}F_k \end{aligned}$$

Como las componentes de las variables en el espacio de los factores comunes vienen dadas por:

$$X_1' \rightarrow (l_{11}, l_{12}, \dots, l_{1k})$$

$$X_2' \rightarrow (l_{21}, l_{22}, \dots, l_{2k})$$

\vdots

$$X_p' \rightarrow (l_{p1}, l_{p2}, \dots, l_{pk})$$

las componentes del centro de gravedad o centroide son:

$$C = \left(\frac{1}{p} \sum_{j=1}^p l_{j1}, \quad \frac{1}{p} \sum_{j=1}^p l_{j2}, \quad \dots, \quad \frac{1}{p} \sum_{j=1}^p l_{jk} \right)$$

Si exigimos que el primer factor pase por C, el centroide tendrá todas sus componentes nulas excepto la primera, es decir:

$$\sum_{j=1}^p l_{j2} = \dots = \sum_{j=1}^p l_{jk} = 0$$

Entonces, en la restricción ya conocida $r_{hj} = l_{h1}l_{j1} + l_{h2}l_{j2} + \dots + l_{hp}l_{jp}$ se puede sumar en j a ambos lados de la igualdad, para obtener la expresión:

$$\sum_{j=1}^p r_{hj} = l_{h1} \sum_{j=1}^p l_{j1} + l_{h2} \sum_{j=1}^p l_{j2} + \dots + l_{hp} \sum_{j=1}^p l_{jp} = l_{h1} \sum_{j=1}^p l_{j1}$$

Si ahora sumamos en h ambos lados de la igualdad anterior tenemos:

$$T = \sum_{h=1}^p \sum_{j=1}^p r_{hj} = \sum_{h=1}^p l_{h1} \sum_{j=1}^p l_{j1} = \left(\sum_{j=1}^p l_{j1} \right)^2$$

Ya podemos escribir lo siguiente:

$$\sum_{j=1}^p r_{hj} = l_{h1} \sum_{j=1}^p l_{j1} \Rightarrow l_{h1} = \frac{\sum_{j=1}^p r_{hj}}{\sum_{j=1}^p l_{j1}} = \frac{\sum_{j=1}^p r_{hj}}{\sqrt{T}} = \frac{S_h}{\sqrt{T}} \quad h = 1 \dots p$$

Ya hemos obtenido los pesos o saturaciones en el primer factor l_{h1} como el cociente entre la suma de correlaciones en la columna h de LL' entre la suma de todas las correlaciones de LL' .

Considerando ahora las correlaciones $r'_{ij} = r_{ij} - l_{i1} l_{j1}$ con $i,j=1\dots p$, las componentes de las variables en los restantes factores serán:

$$(l_{12}, \dots, l_{1k})$$

$$(l_{22}, \dots, l_{2k})$$

⋮

$$(l_{p2}, \dots, l_{pk})$$

Ahora elegimos el segundo factor de modo que pase por el origen y por C y cambiamos de signo ciertas variables X_i para que no se anulen a la vez todas las sumas:

$$\sum_{j=1}^p l_{j2}, \dots, \sum_{j=1}^p l_{jk}$$

Repetiendo el proceso anterior, se obtienen los pesos o saturaciones en el segundo factor l_{h2} como:

$$l_{h2} = \frac{s_h S_{1h}}{\sqrt{T_1}} \quad h = 1 \dots p$$

Ya hemos obtenido los pesos o saturaciones en el segundo factor l_{h2} como el cociente entre la suma de correlaciones en la columna h de la matriz transformada de LL' mediante $r'_{ij} = r_{ij} - l_{i1} l_{j1}$ con $i,j=1\dots p$, entre la suma de todas las correlaciones de dicha matriz transformada. El término s_h cambia de signo al cociente si ha sido necesario cambiar el signo de X_h para que no se anulasen a la vez todas las sumas:

$$\sum_{j=1}^p l_{j2}, \dots, \sum_{j=1}^p l_{jk}$$

El proceso para obtener los demás factores es exactamente el mismo.

MÉTODO DE LAS COMPONENTES PRINCIPALES PARA OBTENER LOS FACTORES

La teoría de componentes principales estudiada en el Capítulo anterior puede utilizarse para la obtención de los factores en el modelo factorial. Es preciso no confundir la Teoría general de componentes principales con una de sus aplicaciones para la obtención de factores en el modelo factorial, que es precisamente lo que se verá aquí.

En el análisis en componentes principales se dispone de una muestra de tamaño n acerca de p variables X_1, X_2, \dots, X_p (tipificadas o no) inicialmente correlacionadas, para posteriormente obtener a partir de ellas un número $k \leq p$ de variables incorrelacionadas Z_1, Z_2, \dots, Z_k que sean combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad. Tendremos entonces que:

$$\begin{aligned} Z_1 &= u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Z_2 &= u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ &\vdots \\ Z_p &= u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{aligned}$$

Pero este sistema de ecuaciones es reversible, siendo posible expresar las variables X_j en función de las componentes principales Z_j de la siguiente forma:

$$\begin{aligned} X_1 &= u_{11}Z_1 + u_{21}Z_2 + \cdots + u_{p1}Z_p \\ X_2 &= u_{21}Z_1 + u_{22}Z_2 + \cdots + u_{p2}Z_p \\ &\vdots \\ X_p &= u_{1p}Z_1 + u_{2p}Z_2 + \cdots + u_{pp}Z_p \end{aligned}$$

La matriz de coeficientes de este segundo sistema es la matriz transpuesta de la matriz de coeficientes del sistema anterior, pudiendo utilizarse este segundo sistema para la estimación de los factores. El único problema que podría presentarse es que las componentes Z_j no estén tipificadas, condición que sí se ha exigido a los factores. Este problema se salva utilizando componentes principales tipificadas, definidas por:

$$Y_j = \frac{Z_j}{\sqrt{\lambda_j}} \quad j=1,2,\dots,p$$

Entonces, en el segundo sistema sustituimos los Z_j por $Y_j \sqrt{\lambda_j}$, resultando la ecuación j -ésima del sistema de la siguiente forma:

$$X_j = u_{1j}Y_1 \sqrt{\lambda_1} + u_{2j}Y_2 \sqrt{\lambda_2} + \cdots + u_{pj}Y_p \sqrt{\lambda_p}$$

Pero, de la Teoría de componentes principales sabemos que $u_{hj} \sqrt{\lambda_h}$ es el coeficiente de correlación entre la variable j -ésima y la componente h -ésima, lo que permite escribir la ecuación como:

$$X_j = r_{1j}Y_1 + r_{2j}Y_2 + \cdots + r_{pj}Y_p$$

pudiéndose separar en esta última ecuación sus últimos $p-k$ términos, lo que permite escribirla como:

$$X_j = r_{1j}Y_1 + r_{2j}Y_2 + \dots + r_{kj}Y_k + (r_{k+1,j}Y_{k+1} + \dots + r_{pj}Y_p)$$

Comparando esta ecuación con la ecuación del modelo factorial:

$$X_j = l_{j1}F_1 + l_{j2}F_2 + \dots + l_{jk}F_k + e_j$$

se observa que los k factores F_h se estiman mediante las k primeras componentes principales tipificadas Y_h y la estimación de los coeficientes l_{jh} viene dada por:

$$\hat{l}_{j1} = r_{1j}, \quad \hat{l}_{j2} = r_{2j}, \quad \dots, \quad \hat{l}_{jk} = r_{kj}$$

pudiéndose estimar la communalidad de la variable X_j como:

$$\hat{h}_j^2 = \hat{l}_{j1}^2 + \hat{l}_{j2}^2 + \dots + \hat{l}_{jk}^2$$

y el factor único e_j se estimará como:

$$\hat{e}_j = r_{k+1,j}Y_{k+1} + r_{m+2,j}Y_{k+2} + \dots + r_{pj}Y_p$$

y la especificidad o parte de la varianza debida al factor único se estima como:

$$\hat{\sigma}_j^2 = 1 - \hat{h}_j^2$$

MÉTODO DE COMPONENTES PRINCIPALES ITERADAS O EJES PRINCIPALES PARA OBTENER LOS FACTORES

Es un método similar al de las componentes principales. Se trata de un método iterativo que comienza con el cálculo de la matriz de correlación muestral:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

A continuación se realiza una estimación inicial de las communalidades de cada variable calculando la regresión de cada variable sobre el resto de variables originales, estimándose la communalidad de la variable mediante el coeficiente de determinación obtenido en la regresión.

El siguiente paso es sustituir en la matriz R cada 1 de la diagonal principal por la estimación de la communalidad correspondiente a cada variable. A la matriz R modificada de esta forma la denominamos matriz de correlación reducida R*:

$$R^* = \begin{bmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{bmatrix} = R - \hat{\Omega}$$

A continuación se calculan las raíces características y los vectores característicos asociados a la matriz R*, a partir de los cuales se obtienen las cargas factoriales estimadas $\hat{\lambda}_{jh}$.

Se determinan los factores a retener k mediante un contraste de este tipo de componentes principales de los ya vistos y se calcula la communalidad de cada variable con los k factores retenidos:

$$\hat{h}_j^2 = \hat{l}_{j1}^2 + \hat{l}_{j2}^2 + \cdots + \hat{l}_{jk}^2$$

y la especificidad o parte de la varianza debida al factor único se estima como:

$$\hat{\omega}_j^2 = 1 - \hat{h}_j^2 \quad j=1, \dots, p$$

Hay que tener presente que todas estas especificidades han de resultar positivas, pues se trata de varianzas. Si alguna resulta negativa puede ser que el método no sea aplicable.

Supuestas todas la especificidades positivas se itera el proceso partiendo de la nueva matriz R* cuya diagonal presenta las communalidades recién estimadas. El procedimiento iterativo se detiene cuando la diferencia entre la communalidad estimada para cada variable entre dos iteraciones sucesivas sea menor que una cantidad prefijada.

MÉTODO DE MÁXIMA VERO SIMILITUD PARA OBTENER LOS FACTORES

Los métodos vistos hasta ahora para obtener los factores pueden considerarse métodos directos, mientras que los métodos que se van a ver a continuación son métodos estrictamente estadísticos basados en la teoría de la inferencia.

Para poder utilizar máxima verosimilitud necesitamos añadir la hipótesis suplementaria de que los vectores $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ constituyen una realización de una muestra aleatoria simple de una población normal p-variante $N_p(\vec{\mu}, \Sigma)$ con $\Sigma = LL' + \Omega$.

Si sustituimos $\vec{\mu}$ por su estimador insesgado $\hat{\vec{\mu}} = \frac{1}{n} \sum_{i=1}^n \vec{x}_i$, la función de verosimilitud toma la siguiente forma:

$$L(\Sigma, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n) = \frac{1}{(2\pi)^{\frac{np}{2}} |\Sigma|^{\frac{n}{2}}} e^{-\frac{n}{2} \text{traza}(\Sigma^{-1}S)}$$

$|\Sigma|$ es el determinante de Σ , y $S = (S_{ij})$ es la matriz de covarianzas muestrales, siendo:

$$S_{ij} = \frac{1}{n} \sum_{s=1}^n (x_{si} - \hat{\mu}_i)(x_{sj} - \hat{\mu}_j) \quad i, j = 1 \dots p$$

Tomando logaritmos en la función de verosimilitud tenemos:

$$\ln[L(\Sigma, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)] = -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln(|\Sigma|) - \frac{n}{2} \text{traza}(\Sigma^{-1}S)$$

Suponiendo que $\Sigma = LL' + \Omega$, la verosimilitud será una función de L y Ω que podremos maximizarla suponiendo además que $L'\Omega^{-1}L$ es diagonal.

De la expresión del logaritmo de la verosimilitud se deduce que su maximización es equivalente a la minimización de la expresión:

$$\ln(|\Sigma|) + \text{traza}(\Sigma^{-1}S)$$

que también es equivalente a la minimización de la expresión:

$$\ln(|\Sigma|) + \text{traza}(\Sigma^{-1}S) - \ln(|S|) - p \quad (\text{la constante adicional no interviene al minimizar})$$

y como $\ln(|\Sigma|) - \ln(|S|) = -\ln(\Sigma^{-1}S)$, la función a minimizar será:

$$f(F, \Omega) = \text{traza}(\Sigma^{-1}S) - \ln(\Sigma^{-1}S) - p \quad \text{donde } \Sigma = LL' + \Omega$$

Para hallar \hat{L} y $\hat{\Omega}$ que minimicen $f(L, \Omega) = \text{traza}(\Sigma^{-1}S) - \ln(\Sigma^{-1}S) - p$ deben igualarse a cero las derivadas parciales de f respecto de L y Ω respectivamente. Lawley y Maxwell demostraron que estas derivadas parciales toman las expresiones:

$$\frac{\partial f(L, \Omega)}{\partial L} = 2\Sigma^{-1}(\Sigma - S)\Sigma^{-1}L$$

$$\frac{\partial f(F, \Omega)}{\partial \Omega} = \text{diag}(\Sigma^{-1}(\Sigma - S)\Sigma^{-1})$$

Luego, las ecuaciones a resolver para hallar \hat{L} y $\hat{\Omega}$ que minimicen f serán:

$$\begin{cases} \Sigma^{-1}(\Sigma - S)\Sigma^{-1}L = 0 \\ \text{diag}(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}) = \text{diag}(\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}) = 0 \\ \Sigma = LL' + \Omega \\ J = L'\Omega^{-1}L \quad \text{diagonal} \end{cases}$$

Las ecuaciones anteriores determinan \hat{L} y $\hat{\Omega}$ sólo de forma implícita. Las soluciones explícitas requieren la utilización de métodos iterativos de cálculo numérico.

Lawley demostró que las ecuaciones anteriores son equivalentes a:

$$\begin{cases} L = (S - \Omega)\Omega^{-1}LJ^{-1} \\ \text{diag}(\Sigma - S) = 0 \Rightarrow v_i = s_{ii} - \sum_{j=1}^k a_{ij}^2 \end{cases}$$

En la primera de las dos ecuaciones anteriores se observa que los elementos de la diagonal de J son los valores propios de $\Omega^{-1}(S - \Omega)$, y las columnas de la matriz factorial L son los correspondientes valores propios, con lo cual ya hemos determinado \hat{L} . También se demuestra que $\hat{V} = \text{diag}(S - \hat{L}\hat{L}')$.

Si se trabaja con variables reducidas, el proceso es similar, sustituyendo la matriz S por la matriz R de correlaciones muestrales. La solución de Lawley sólo tiene como exigencia que Ω^{-1} exista.

Por otra parte, **Joreskog** demostró que las ecuaciones del proceso de minimización de f para una Ω dada son equivalentes a:

$$(\Omega^{-1/2} S \Omega^{-1/2})(\Omega^{-1/2} L) = (\Omega^{-1/2} L)(I + J)$$

lo que nos lleva a la conclusión de que $\Omega^{-1/2} L$ son vectores propios de $\Omega^{-1/2} S \Omega^{-1/2}$ relativos a los valores propios de los elementos de la diagonal de $I+J$. Entonces, si $\hat{\Theta}$ es la matriz diagonal de orden k con los k primeros valores propios en orden creciente, y \hat{W} es la matriz cuyas columnas son los vectores propios, entonces se demuestra que, para Ω dada, la estimación de la matriz factorial L resulta ser:

$$\hat{L} = \Omega^{-1/2} \hat{W} (\hat{\Theta} - I)^{-1/2}$$

Por otro lado, las ecuaciones del proceso de minimización de f para una L dada son equivalentes a:

$$\frac{\partial f}{\partial \Omega} = \text{diag}(\Omega^{-1} (\hat{L} \hat{L}' + \Omega - S) \Omega^{-1}) = 0 \Rightarrow \hat{\Omega} = \text{diag}(S - \hat{L} \hat{L}')$$

La solución de Joreskog sólo tiene como exigencia que $\Omega^{-1/2}$ exista. Si alguna unicidad es prácticamente nula, hay problemas con la existencia de $\Omega^{-1/2}$, pero en este caso se obtiene la solución por componentes principales para estas variables con unicidades prácticamente nulas, analizando a continuación las demás variables por el método de la máxima verosimilitud, y combinando finalmente ambos métodos para dar una solución completa al conjunto de todas las variables.

MÉTODOS MINRES, ULS Y GLS PARA OBTENER LOS FACTORES

El método MINRES (*Minimizing residuals*) también denominado ***análisis factorial de correlaciones*** calcula la matriz factorial $L = (l_{ij})$ que minimiza los residuos:

$$\bar{r}_{hs} = r_{hs} - \sum_{t=1}^k l_{ht} l_{jt} \quad h \neq s$$

Por lo tanto se calcula la matriz factorial minimizando las diferencias entre la correlación observada y la deducida del modelo factorial, excepto para las correlaciones unitarias $r_{ii}=1$. El criterio de minimización a utilizar es el de los mínimos cuadrados, tratando de hallar L que haga mínima la suma de cuadrados de los elementos no diagonales de la matriz residual $\bar{R} = R * -LL'$, o sea, que haga mínima la función:

$$F(L) = \sum_{j=h+1}^p \sum_{h=1}^{p-1} \left(r_{hj} - \sum_{t=1}^k a_{ht} a_{jt} \right)^2$$

Este método permite estimar L sin necesidad de determinar previamente las comunidades, que a su vez se obtienen como resultado del método. Además, en este método no es necesario suponer hipótesis alguna de multinormalidad de las variables, como ocurre en el caso del método de máxima verosimilitud.

La estimación de L a partir de la minimización de $F(L)$ puede dar lugar a comunidades mayores que la unidad, hecho que se evitará imponiendo a $F(L)$ la restricción:

$$h_j^2 = \sum_{t=1}^k l_{jt}^2 \leq 1 \quad j = 1 \dots n$$

Para encontrar una solución efectiva de L , lo más adecuado es utilizar el procedimiento iterativo de Gauss-Seidel de resolución de sistemas de ecuaciones lineales.

Joreskog enfoca de forma general la estimación de la matriz factorial estableciendo una relación entre el método de máxima verosimilitud y los métodos basados en el criterio de los mínimos cuadrados. La estimación de L tal que $\Sigma = LL' + \Omega$ obtenida a partir de la matriz de covarianzas muestrales S , para un k dado, se puede conseguir de tres formas distintas:

1º) **Método de los mínimos cuadrados no ponderados ULS (unweighted least squares)**, que consiste en minimizar la función:

$$U(L, \Omega) = \text{traza}(S - \Sigma)^2 / 2$$

2º) **Método de los mínimos cuadrados generalizados GLS (generalized least squares)**, que consiste en minimizar la función:

$$G(L, \Omega) = \text{traza}(I - S^{-1}\Sigma) / 2$$

3º) **Método de máxima verosimilitud ML (maximum likelihood)**, que consiste en minimizar la función:

$$F(L, \Omega) = \log(\Sigma) + \text{traza}(S\Sigma^{-1}) - \log(|S|) - p$$

Las tres funciones pueden ser minimizadas mediante el mismo método básico consistente en dos pasos. En el primer paso, se halla el mínimo condicional para Ω dado, lo que produce como resultado una función $f(\Omega)$. En el segundo paso, se minimiza en Ω esta función mediante un método numérico, que generalmente suele ser el de Newton-Raphson, en caso de que se puedan hallar las dos primeras derivadas de la función. Joreskog ofrece las soluciones en función de los valores y vectores propios de la matriz $S - \Omega$ en el método ULS, y de la matriz $\Omega S^{-1} \Omega$ en los métodos GLS y ML.

Joreskog interpreta la solución por el método del factor principal y la solución MINRES como equivalentes al método ULS, y afirma que los métodos GLS y ML son invariantes para cambio de escala. Posteriormente se demostró mediante contraejemplos que los métodos ML y GLS no siempre son invariantes, mientras que el método ULS puede serlo en determinadas circunstancias.

CONTRASTES EN EL MODELO FACTORIAL

En el modelo factorial pueden realizarse varios tipos de contrastes. Estos contrastes suelen agruparse en dos bloques, según se apliquen previamente a la extracción de los factores o que se apliquen después. Con los contrastes aplicados previamente a la extracción de los factores trata de analizarse la pertinencia de la aplicación del análisis factorial a un conjunto de variables observables. Con los contrastes aplicados después de la obtención de los factores se pretende evaluar el modelo factorial una vez estimado.

Dentro del grupo de *contrastos que se aplican previamente a la extracción de los factores* tenemos el contraste de esfericidad de Barlett y la medida de adecuación muestral de Kaiser, Meyer y Olkin.

Contraste de esfericidad de Barlett

Evidentemente, antes de realizar un análisis factorial nos plantearemos si las p variables originales están correlacionadas entre sí o no lo están. Si no lo estuvieran no existirían factores comunes y, por lo tanto, no tendría sentido aplicar el análisis factorial. Esta cuestión suele probarse utilizando el contraste de esfericidad de Barlett.

La matriz de correlación poblacional R_p recoge la relación entre cada par de variables mediante sus elementos ρ_{ij} situados fuera de la diagonal principal. Los elementos de la diagonal principal son unos, ya que toda variable está totalmente relacionada consigo misma. En caso de que no existiese ninguna relación entre las p variables en estudio, la matriz R_p sería la identidad, cuyo determinante es la unidad. Por lo tanto, para decidir la ausencia o no de relación entre las p variables puede plantearse el siguiente contraste:

$$\begin{aligned} H_0 &: |R_p|=1 \\ H_1 &: |R_p|\neq 1 \end{aligned}$$

Barlett introdujo un estadístico para este contraste basado en la matriz de correlación muestral R , que bajo la hipótesis H_0 tiene una distribución chi-cuadrado con $p(p-1)/2$ grados de libertad. La expresión de este estadístico es la siguiente:

$$-[n-1-(2p+5)/6]\ln|R|$$

Medida KMO de Kaiser, Meyer y Olkin de adecuación muestral global al modelo factorial y medida MSA de adecuación individual

En un modelo con varias variables el coeficiente de correlación parcial entre dos variables mide la correlación existente entre ellas una vez que se han descontado los efectos lineales del resto de las variables del modelo. En el modelo factorial se pueden considerar esos efectos de otras variables como los correspondientes a los factores comunes. Por lo tanto, el coeficiente de correlación parcial entre dos variables sería equivalente al coeficiente de correlación entre los factores únicos de esas dos variables. Pero de acuerdo con el modelo de análisis factorial los coeficientes de correlación teóricos calculados entre cada par de factores únicos son nulos por hipótesis, y como los coeficientes de correlación parcial constituyen una aproximación a dichos coeficientes teóricos, deben estar próximos a cero. Kaiser-Meyer y Olkin definen la medida KMO de adecuación muestral global al modelo factorial basada en los coeficientes de correlación observados de cada par de variables y en sus coeficientes de correlación parcial mediante la expresión siguiente:

$$KMO = \frac{\sum_{j} \sum_{h \neq j} r_{jh}^2}{\sum_{j} \sum_{h \neq j} r_{jh}^2 + \sum_{j} \sum_{h \neq j} a_{jh}^2}$$

r_{jh} son los coeficientes de correlación observados entre las variables X_j y X_h
 a_{jh} son los coeficientes de correlación parcial entre las variables X_j y X_h

En el caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador, que recoge los coeficientes a_{jh} , será pequeño y, en consecuencia, la medida KMO será próxima a la unidad. Valores de KMO por debajo de 0,5 no serán aceptables, considerándose inadecuados los datos a un modelo de análisis factorial. Para valores superiores a 0,5 se considera aceptable la adecuación de los datos a un modelo de análisis factorial. Mientras más cerca estén de 1 los valores de KMO mejor es la adecuación de los datos a un modelo factorial, considerándose ya excelente la adecuación para valores de KMO próximos a 0,9.

También existe una medida de adecuación muestral individual para cada una de las variables basada en la medida KMO. Esta medida se denomina MSA (*Measure of Sampling Adequacy*), se define de la siguiente forma:

$$MSA_j = \frac{\sum_{h \neq j} r_{jh}^2}{\sum_{h \neq j} r_{jh}^2 + \sum_{h \neq j} a_{jh}^2}$$

Si el valor de MSA_j se aproxima a la unidad, la variable X_j será adecuada para su tratamiento en el análisis factorial con el resto de las variables.

También en el modelo factorial pueden realizarse *contrastes después de la obtención de los factores con los que se pretende evaluar el modelo factorial una vez estimado*. Entre ellos tenemos el contraste para la bondad de ajuste del método de máxima verosimilitud y el contraste para la bondad de ajuste del método MINRES.

Contraste para la bondad de ajuste en el método ML de máxima verosimilitud

Una de las principales ventajas de la estimación del modelo factorial por máxima verosimilitud es que proporciona un contraste para la hipótesis:

H_0 : k factores son suficientes para describir los datos.

H_1 : La matriz Σ no tiene restricciones.

Sabemos que bajo H_0 la función de máxima verosimilitud es:

$$L_0(\hat{\Sigma}, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n) = \frac{1}{(2\pi)^{\frac{np}{2}} |\hat{\Sigma}|^{\frac{n}{2}}} e^{-\frac{n}{2} \text{traza}(\hat{\Sigma}^{-1} S)}$$

con $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Omega}$, siendo \hat{L} y $\hat{\Omega}$ los estimadores de máxima verosimilitud obtenidos. Bajo H_1 sabemos que el estimador de máxima verosimilitud de Σ es S , en cuyo caso la función de verosimilitud será:

$$L_1(\hat{\Sigma}, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n) = \frac{1}{(2\pi)^{\frac{np}{2}} |S|^{\frac{n}{2}}} e^{-\frac{n}{2} \text{traza}(I)} = \frac{1}{(2\pi)^{\frac{np}{2}} |S|^{\frac{n}{2}}} e^{-\frac{n}{2} p}$$

Si llamamos $\lambda = \frac{L_0}{L_1}$ sabemos que el contraste de razón de verosimilitud se realiza utilizando el estadístico:

$$\begin{aligned}
 -2\ln\lambda &= -2\left[\frac{n}{2}\ln(|S|) + \frac{n}{2}p - \frac{n}{2}\ln\left(\hat{\Sigma}\right) - \frac{n}{2}\text{traza}(\hat{\Sigma}^{-1}S)\right] = \\
 -np - n\ln\left(\frac{|S|}{|\hat{\Sigma}|}\right) + n\text{traza}(\hat{\Sigma}^{-1}S) &= np\left[\frac{1}{p}\text{traza}(\hat{\Sigma}^{-1}S) - \frac{1}{p}\ln\left(|\hat{\Sigma}^{-1}S|\right) - 1\right] = \\
 np(\hat{a} - \ln(\hat{g}) - 1)
 \end{aligned}$$

donde \hat{a} y \hat{g} son, respectivamente, las medias aritmética y geométrica de los valores propios de $\hat{\Sigma}^{-1}S$.

Sabemos que, si H_0 es cierta, $-2\ln\lambda$ sigue asintóticamente una distribución chi-cuadrado con s grados de libertad, donde:

$$s = p(p+1)/2 - [p(k+1) - k(k-1)/2] = (p-k)^2 - (p+k)$$

luego para una probabilidad de error de tipo I de valor α , la región de aceptación de H_0 es:

$$\{np(\hat{a} - \ln(\hat{g}) - 1) \leq \chi^2_{s,\alpha}\}$$

Bartlett demostró que la aproximación a la chi-cuadrado mejoraba si se sustituía n por $n' = n - (2p+5)/5 - 2k/3$.

En el caso trivial, $k=0$, H_0 es la hipótesis de que las variables son independientes. En este caso el estimador de máxima verosimilitud de Σ es $\hat{\Sigma} = \text{diag}(S)$, y como:

$$\text{traza}(\hat{\Sigma}^{-1}S) = \text{traza}(\hat{\Sigma}^{-1/2}S\hat{\Sigma}^{-1/2}) = \text{traza}(R) = p$$

donde R es la matriz de correlaciones de los datos.

El estadístico del contraste será $-n \ln(|R|)$.

En la práctica, el problema es decidir cuántos factores comunes es razonable que se ajusten a los datos. Para hacer esto se sigue un procedimiento secuencial partiendo de un valor pequeño para k ($k=0$ o $k=1$) y se aumenta el número de factores comunes de uno en uno hasta que no se rechace H_0 . Este procedimiento, sin embargo, ha sido objeto de críticas, ya que los valores críticos del contraste no se ajustan para tener en cuenta que se contrastan secuencialmente un conjunto de hipótesis.

Para determinados datos, el modelo factorial se rechaza para todos los valores de k para los que $s>0$. En tales casos concluimos que no existe modelo factorial que se ajuste a los datos.

Contraste para la bondad de ajuste en el método MINRES

La estimación MINRES también proporciona un contraste para las hipótesis:

H_0 : k factores son suficientes para describir los datos.

H_1 : La matriz Σ no tiene restricciones.

El contraste se basa en el estadístico $U_k=(N-1)\log(|LL'+\Omega| / |R|)$ cuya distribución asintótica bajo la hipótesis nula H_0 y suponiendo multinormalidad de las variables es una chi-cuadrado con $d = [(p-k)^2 - p - k]/2$ grados de libertad.

Este estadístico suele aparecer también bajo la forma:

$$U_k=N[\log(|LL'+\Omega|) - \log(|R|) + \text{traza}(R(LL'+\Omega)^{-1}) - p]$$

e incluso, siendo más sofisticados, podría utilizarse el valor $N-1-(2p+5)/5-2m/3$ en lugar de N , sobre todo cuando N no es muy grande.

INTERPRETACIÓN GEOMÉTRICA DEL ANÁLISIS FACTORIAL

En el análisis factorial puede realizarse una representación geométrica de las variables aleatorias objeto del análisis. Podemos representar una variable X por un vector considerando la desviación típica $\sigma(X)$ como la norma (módulo o longitud) del mismo ($\|X\|=\sigma(X)$). Esto es lógico hacerlo ya que una variable es tanto más variable cuanto más dispersa está respecto de su media (representa una variabilidad mayor), y la medida de esta dispersión es precisamente la desviación típica.

Pero para representar completamente la variable X por un vector tenemos que asignarle una dirección y un sentido. La dirección de un vector se determina con referencia a otro vector, que se supone fijo, en función del coseno del ángulo que forman ($-1 \leq \cos(\phi) \leq +1$), que vale $+1$ si ambos vectores tienen la misma dirección y sentido, que vale -1 si ambos vectores tienen la misma dirección y distinto sentido, y que vale cero en caso de que ambos vectores sean ortogonales. Para dos de nuestras variables X e Y podemos considerar su coeficiente de correlación ρ que también verifica $-1 \leq \rho \leq +1$, y que suponiendo que X e Y tienen media nula, $\rho=\pm 1$ implica $Y=aX$. Luego $\rho=+1$ implica que las dos variables tengan la misma dirección y sentido, y $\rho=-1$ implica que las dos variables tengan la misma dirección y distinto sentido.

Intuitivamente podemos decir que dos variables están separadas al máximo si son incorrelacionadas, y como la separación máxima en cuanto a la dirección de dos factores es la ortogonalidad, dicha separación referida a variables será la incorrelación. Podemos entonces poner $\rho = \cos(\varphi)$ y utilizar la correlación para dar una idea de la separación direccional de dos variables aleatorias consideradas como vectores. La separación máxima.

Por lo tanto, podemos representar cada variable del análisis factorial por un vector de módulo igual a la desviación típica de la variable y con origen en el origen de coordenadas. El coseno del ángulo formado por dos vectores es el coeficiente de correlación de las correspondientes variables, correspondiéndose la incorrelación entre dos variables con la ortogonalidad entre dos vectores.

Si las variables son reducidas estarán representadas por vectores de módulo unidad. Si las variables son linealmente independientes (rango n), estos vectores ocuparán un espacio n -dimensional con sus extremos en una esfera de radio unidad, siendo los factores vectores unitarios ortogonales y ortonormales.

En el modelo factorial tenemos:

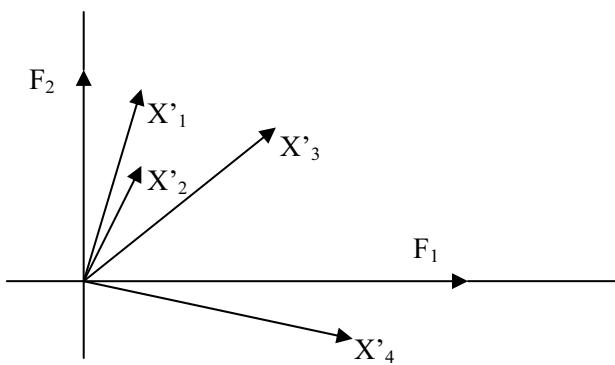
$$\begin{aligned} X_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1k}F_k + e_1 \\ X_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2k}F_k + e_2 \\ &\vdots \\ X_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pk}F_k + e_k \end{aligned}$$

con lo que podemos representar las variables del modelo una vez restadas sus unicidades en función de k factores como sigue:

$$\begin{aligned} X_1' &= X_1 - e_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1k}F_k \\ X_2' &= X_2 - e_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2k}F_k \\ &\vdots \\ X_p' &= X_p - e_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pk}F_k \end{aligned}$$

De esta forma tenemos las variables del análisis factorial sin sus unicidades representadas por los factores, siendo el módulo de estas variables X_j' inferior a la unidad (es exactamente h , pues h_j^2 es la parte de la varianza de la variable X_j debida a los factores comunes, es decir, $V(X_j')$, es decir, la communalidad). Las saturaciones, pesos o cargas factoriales de cada variable en cada factor se representarán por las proyecciones ortogonales de cada variable en cada factor.

A continuación se presenta un gráfico relativo a cuatro variables X_1, X_2, X_3 y X_4 representadas por dos factores F_1 y F_2 .



Como las saturaciones, pesos o cargas factoriales de cada variable en cada factor (elementos de la matriz factorial), se representan por las proyecciones ortogonales de cada variable en cada factor, la cuarta variable se explica fuertemente y de forma positiva por el primer factor (proyección positiva grande de X'_4 sobre F_1), mientras que se representa poco y en sentido negativo por el segundo factor (proyección negativa pequeña de X'_4 sobre F_2). De la misma forma, la primera y segunda variables se explican fuertemente y de forma positiva por el segundo factor, y se explican poco y de forma positiva por el primer factor. La tercera variable se explica de igual forma por el primero y segundo factor.

Puede ocurrir que al realizar esta representación geométrica del modelo factorial, las proyecciones de la mayoría de las variables sobre los factores no sean lo suficientemente grandes como para que la interpretación del modelo resulte adecuada. Si la representación geométrica resulta difusa, se puede realizar una rotación de los factores que clarifique las proyecciones de las variables sobre ellos. Nos introducimos así en el campo de las rotaciones factoriales, que se explicarán con más detalle en los siguientes apartados.

Con una rotación factorial se transforma una solución factorial inicial en otro tipo de solución preferida. Tal transformación va encaminada a poner de manifiesto la solución de la manera más convincente y clara para su interpretación científica.

Teóricamente puede justificarse la interpretación vectorial realizada aquí por el hecho de que el conjunto de variables aleatorias sobre una población dotado de las operaciones de suma y producto por un escalar, tiene estructura de espacio vectorial. Las variables con varianza finita y esperanza nula forman un subespacio vectorial del anterior en el cual la covarianza $\text{Cov}(X, Y)$ es un producto escalar que define una norma dada por la varianza y un ángulo entre vectores cuyo coseno es el coeficiente de correlación.

ROTACIÓN DE LOS FACTORES

El trabajo en el análisis factorial persigue que los factores comunes tengan una interpretación clara, porque de esa forma se analizan mejor las interrelaciones existentes entre las variables originales. Sin embargo, en muy pocas ocasiones resulta fácil encontrar una interpretación adecuada de los factores, iniciales, con independencia del método que se haya utilizado para su extracción. Precisamente los procedimientos de *rotación de factores* se han ideado para obtener, a partir de la solución inicial, unos factores que sean fácilmente interpretables.

En la solución inicial cada uno de los factores comunes están correlacionados en mayor o menor medida con cada una de las variables originales. Pues bien, con los *factores rotados* se trata de que cada una de las variables originales tenga una correlación lo más próxima a 1 que sea posible con uno de los factores y correlaciones próximas a 0 con el resto de los factores. De esta forma, y dado que hay más variables que factores comunes, cada factor tendrá una correlación alta con un grupo de variables y baja con el resto de variables. Examinando las características de las variables de un grupo asociado a un determinado factor se pueden encontrar rasgos comunes que permitan identificar el factor y darle una denominación que responda a esos rasgos comunes. Si se consigue identificar claramente estos rasgos, se habrá dado un paso importante, ya que con los factores comunes no sólo se reducirá la dimensionalidad del problema, sino que también se conseguirá desvelar la naturaleza de las interrelaciones existentes entre las variables originales.

Existen dos formas básicas de realizar la rotación de factores: rotación ortogonal y rotación oblicua. En la *rotación ortogonal*, los ejes se rotan de forma que quede preservada la incorrelación entre los factores. Dicho de otra forma, los nuevos ejes, o ejes rotados, son perpendiculares de igual forma que lo son los factores sin rotar. Por esta restricción, a la rotación ortogonal se le denomina también *rotación rígida*. Entre los diversos procedimientos de rotación ortogonal el denominado método *Varimax* es el más conocido y aplicado. Los ejes de los factores del método *Varimax* se obtienen maximizando la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor. Existen otros métodos de rotación ortogonal de los factores menos utilizados, como son el método *Equamax* y el método *Quartimax*.

En la *rotación oblicua* los ejes no son ortogonales y los factores ya no estarán incorrelacionados, con lo que se pierde una propiedad que en principio es deseable que cumplan los factores. Sin embargo, en ocasiones puede compensarse esta pérdida, si, a cambio, se consigue una asociación más nítida de cada una de las variables con el factor correspondiente. El método de rotación oblicua más conocido es el denominado *Oblimin*, existiendo otros menos utilizados como el *Oblimax*, *Promax*, *Quartimin*, *Biquartimin* y *Covarimin*, algoritmos que permiten controlar el grado de no ortogonalidad. Conviene advertir que tanto en la rotación ortogonal, como en la rotación oblicua la communalidad de cada variable no se ve modificada.

La obtención de la matriz factorial es en general el primer paso de la factorización. El siguiente paso es la rotación de los factores a fin de obtener unos nuevos factores que tengan mayor interpretabilidad. Los diferentes criterios de rotación se rigen por el **postulado de parsimonia**, mediante el cual se elegirá el número mínimo de factores comunes compatible con las variables, y entre las diferentes clases de factores, se elegirán aquellos cuya estructura goce de mayor simplicidad.

ROTACIONES ORTOGONALES

En la rotación ortogonal se plantea el problema siguiente: dada la matriz factorial L, hallar una matriz ortogonal de transformación T, de modo que la matriz $B=LT$ sea la matriz factorial de unos nuevos factores ortogonales, verificando ciertas condiciones analíticas de estructura simple definidas por los distintos métodos de rotación.

Método Varimax

El método *Varimax* obtiene los ejes de los factores maximizando la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor. Suele definirse la *simplicidad* de un factor por la varianza de los cuadrados de sus cargas factoriales en las variables observables. La simplicidad S_i^2 del factor F_i será entonces:

$$S_i^2 = \frac{1}{p} \sum_{j=1}^p (l_{ji}^2)^2 - \left(\frac{1}{p} \sum_{j=1}^p l_{ji}^2 \right)^2$$

El método de rotación Varimax pretende hallar $B=LT$ de modo que la suma de las simplicidades de todos los factores sea máxima, lo que implica la maximización de:

$$S^2 = \sum_{i=1}^k S_i^2 = \sum_{i=1}^k \left[\frac{1}{p} \sum_{j=1}^p (l_{ji}^2)^2 - \left(\frac{1}{p} \sum_{j=1}^p l_{ji}^2 \right)^2 \right]$$

El problema que plantea la expresión anterior es que las variables con mayores communalidades tienen una mayor influencia en la solución final. Para solventar este problema se efectúa la normalización de Kaiser, en la que cada carga factorial al cuadrado se divide por la communalidad de la variable correspondiente (*método Varimax normalizado*). La función a maximizar será ahora:

$$SN^2 = \sum_{i=1}^k \left[\frac{1}{p} \sum_{j=1}^p \left(\frac{l_{ji}^2}{h_j^2} \right)^2 - \left(\frac{1}{p} \sum_{j=1}^p \frac{l_{ji}^2}{h_j^2} \right)^2 \right]$$

En su forma definitiva, el **método Varimax** halla la matriz B maximizando:

$$W = p^2 SN^2 = p \sum_{i=1}^k \sum_{j=1}^p \left(\frac{l_{ji}^2}{h_j^2} \right)^2 - \sum_{i=1}^n \left(\sum_{j=1}^p \frac{l_{ji}^2}{h_j^2} \right)^2$$

Para realizar la maximización se halla la matriz $T = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix}$

que efectúa la rotación de dos factores de forma que su suma de simplicidades sea máxima. Repitiendo esto para los $p(p-1)/2$ pares posibles de factores, se tiene:

$$B = LT_{11}T_{12}T_{13}\dots T_{m-1,m}$$

Cuando la rotación es de más de dos factores se realiza un procedimiento iterativo. El primer y segundo factor se giran según el ángulo φ determinado por el procedimiento anterior. El nuevo primer factor se gira con el tercer factor, y se sigue así hasta que todos los $k(k-1)/2$ pares de factores hayan sido girados. Esta sucesión de rotaciones se llama ciclo. Se repiten los ciclos hasta completar uno en que todos los ángulos de giro sean menores que un cierto valor prefijado.

Una propiedad importante del método *Varimax* es que, después de aplicado, queda inalterada, tanto la varianza total explicada por los factores, como la communalidad de cada una de las variables. La nueva matriz corresponde también a factores ortogonales y tiende a simplificar la matriz factorial por columnas, siendo muy adecuada cuando el número de factores es pequeño.

Método Quartimax

Cuando se realizan rotaciones de los factores se maximizan unas cargas factoriales a costa de minimizar otras. En el caso de la rotación *Quartimax* se hace máxima la suma de las cuartas potencias de todas las cargas factoriales, esto es:

$$Q = \sum_{j=1}^p \sum_{i=1}^k (l_{ji}^2)^2 = \sum_{j=1}^p \sum_{i=1}^k l_{ji}^4 \text{ debe ser máximo}$$

Si T es la matriz ortogonal de la transformación y B=LT, las comunidades $\sum_{i=1}^k b_{ji}^2 = \sum_{i=1}^k l_{ji}^2 = h_j^2$ permanecen invariantes, con lo que también permanecerá constante su cuadrado $\left(\sum_{i=1}^k b_{ji}^2\right)^2 = \sum_{i=1}^k b_{ji}^4 + 2\sum_{i<r}^k b_{ji}^2 b_{jr}^2$. Sumando las p variables se tiene:

$$\sum_{j=1}^p \sum_{i=1}^k b_{ji}^4 + 2\sum_{j=1}^p \sum_{i<r}^k b_{ji}^2 b_{jr}^2 = \text{constante}$$

En esta expresión, el término de la izquierda es Q, y su maximización implica la minimización del término de la derecha $N = \sum_{j=1}^p \sum_{i<r}^k b_{ji}^2 b_{jr}^2$, lo que da una estructura más simple a la matriz B.

También puede considerarse la varianza de los cuadrados de todas las cargas factoriales de la matriz, obteniéndose la expresión:

$$M = \frac{1}{kp} \sum_{j=1}^p \sum_{i=1}^k b_{ji}^4 - (\bar{b}^2)^2 \quad \bar{b}^2 = \frac{1}{kp} \sum_{j=1}^p \sum_{i=1}^k b_{ji}^2$$

La maximización de M también es un buen criterio de estructura simple.

Otro camino distinto consiste en hallar la matriz factorial B de modo que la curtosis de los cuadrados de sus cargas factoriales sea máxima. Tendremos que:

$$K = \frac{\sum_{j=1}^p \sum_{i=1}^k b_{ji}^4}{\left(\sum_{j=1}^p \sum_{i=1}^k b_{ji}^2\right)^2} \quad \text{debe ser máximo}$$

En resumen, hemos planteado cuatro criterios analíticos de estructura simple (Q máximo, N mínimo, K máximo y M máximo) todos ellos equivalentes. La obtención de B que verifique uno cualquiera de los criterios anteriores se consigue maximizando Q. Para obtener la matriz B que maximiza Q se sigue un proceso análogo al de la rotación Varimax.

La nueva matriz corresponde también a factores ortogonales y tiende a simplificar la matriz factorial por filas, siendo muy adecuada cuando el número de factores es elevado.

Métodos Ortomax: Ortomax general, Biquartimax y Equamax

Realmente sólo existen dos métodos distintos para conseguir rotaciones ortogonales que se aproximen a la estructura simple, que son el método *Varimax* y el método *Quartimax*.

El **método Ortomax general** considera una solución intermedia a los métodos *Varimax* y *Quartimax*, maximizando la función:

$$B = \alpha Q + \beta W$$

siendo α y β parámetros a elegir.

Poniendo $v=\beta/(\alpha+\beta)$, y después de algunas transformaciones, el llamado *método Ortomax en su forma general* consiste en maximizar:

$$\sum_{r=1}^k \left(\sum_{j=1}^p b_{jr}^4 - \frac{v}{p} \left(\sum_{j=1}^p b_{jr}^2 \right)^2 \right)$$

Si $v=0$ el *método Ortomax general* equivale al método *Quartimax*, y si $v=1$ equivale al método *Varimax*. Si $v=1/2$ tenemos el **método Biquartimax** o criterio igualmente ponderado. Si $v=k/2$ tenemos el **método Equamax**

ROTACIONES OBLÍCUAS

Dado el modelo factorial L (de factores ortogonales), nos proponemos hallar una matriz T, de modo que $P=LT$ verifique unos criterios de estructura simple. No imponemos ahora restricción a la matriz T, es decir, T puede ser no ortogonal. Esto significa que la matriz P corresponderá a unos factores oblícuos y contendrá las cargas factoriales de las variables en los factores oblícuos. No obstante, existe una matriz V de estructura factorial oblicua tal que $V=PD=L\Lambda$ siendo D diagonal y Λ coincidente con T normalizada por filas, con lo que las columnas de V son las mismas que las de P multiplicadas por una constante. Por lo tanto la optimización de P implica la de V y viceversa.

Existen varios métodos para alcanzar una estructura simple oblicua, unos sobre la matriz V y otros sobre la matriz P.

Método Oblimax y método Quartimin

En el **método Oblimax** se halla Λ de modo que los coeficientes de $V=L\Lambda$ verifiquen que:

$$K = \frac{\sum_{j=1}^p \sum_{i=1}^k v_{ji}^4}{\left(\sum_{j=1}^p \sum_{i=1}^k v_{ji}^2 \right)^2} \text{ sea máximo}$$

Para obtener Λ se empieza rotando un par de factores mediante una matriz cualquiera que maximice K para este par. Esto se repite para todos los pares completando un ciclo, ciclos que también se repiten hasta que K alcance el máximo.

En el **método Quartimin** se minimiza:

$$N = \sum_{j=1}^p \sum_{i < r}^k v_{ji}^2 v_{jr}^2$$

El proceso numérico para hallar Λ exige una larga iteración, en la que cada paso es la obtención de los vectores propios de una matriz simétrica.

Métodos Oblimin: Covarimin, Oblimin general y Biquartimin

Se trata de la adaptación de la rotación *Varimax* al caso oblícuo. Un primer método es el **método Covarimin** que consiste en minimizar las covarianzas de los cuadrados de los coeficientes de V . Es decir, se trata de minimizar la expresión:

$$C = \sum_{r < s = 1}^k \left(p \sum_{j=1}^p v_{jr}^2 v_{js}^2 - \sum_{j=1}^p v_{jr}^2 \sum_{j=1}^p v_{js}^2 \right)$$

Se demuestra que en el caso ortogonal equivale al método *Varimax*. El inconveniente de este método es que proporciona factores casi ortogonales, en contraste con los factores muy oblícuos que proporciona el método *Quartimin*.

El **método Oblimin general** considera una solución intermedia a los métodos *Covarimin* y *Quartimin*, minimizando la función:

$$B = \alpha N + \beta C/n$$

siendo α y β parámetros a elegir.

Poniendo $v = \beta / (\alpha + \beta)$, y después de algunas transformaciones, el llamado *método Oblimin en su forma general* consiste en minimizar:

$$B = \sum_{r < s=1}^k \left(p \sum_{j=1}^p v_{jr}^2 v_{js}^2 - v \sum_{j=1}^p v_{jr}^2 \sum_{j=1}^p v_{js}^2 \right)$$

El grado de oblicuidad de los factores depende del parámetro $0 \leq v \leq 1$. Si $v=0$ equivale al método *Quartimin* (máxima oblicuidad), y si $v=1$ equivale al método *Covarimin* (mínima oblicuidad). Si $v=1/2$ tenemos el **método Biquartimin**.

Método Oblimin directo: Rotación Promax

El **método Oblimin directo** consiste en hallar P de modo que sea mínimo:

$$f(P) = \sum_{r < s=1}^k \left(\sum_{j=1}^p p_{jr}^2 p_{js}^2 - \frac{\delta}{p} \sum_{j=1}^p p_{jr}^2 \sum_{j=1}^p p_{js}^2 \right)$$

siendo δ un parámetro que determina el grado de oblicuidad de los factores.

El **método de Rotación Promax** es un método directo calculable sin necesidad de procesos iterativos, resultando más simple que el resto de los métodos de rotación oblicua. Este método se aplica directamente a la matriz factorial ortogonal rotada según el criterio *Varimax*.

Sea A la matriz factorial ortogonal rotada según el método *Varimax*. Se construye la matriz $P=(P_{ij})$ siendo:

$$P_{ij} = \frac{|a_{ij}^{r+1}|}{a_{ij}} \quad r > 1$$

Cada elemento de P es la potencia r -ésima del respectivo elemento de A conservando el signo. Una carga factorial a_{ij} grande elevada a la potencia r quedará mucho más destacada que una saturación pequeña. A continuación se calcula L tal que AL coincida con P en el sentido de los mínimos cuadrados, siendo la solución:

$$L = (A'A)^{-1} A'P$$

La matriz L debe ser normalizada de modo que $T = (L')^{-1}$ tenga sus factores columna de módulo unidad. Entonces $P = AL$ es el modelo factorial oblicuo y el grado de oblicuidad de los factores obtenidos aumenta con el valor entero r , que juega un papel parecido a v en los métodos *Oblimin*.

PUNTUACIONES O MEDICIÓN DE LOS FACTORES

El análisis factorial es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por los factores obtenidos. Por ejemplo en el caso de estimación de modelos afectados de multicolinealidad. Por ello, es necesario conocer los valores que toman los factores en cada observación. Sin embargo, es importante hacer constar que, salvo el caso de que se haya aplicado el análisis de componentes principales para la extracción de factores, no se obtienen unas puntuaciones exactas para los factores. En su lugar, es preciso realizar estimaciones para obtenerlas. Estas estimaciones se pueden realizar por distintos métodos. Los procedimientos más conocidos, y que aparecen implementados en los paquetes de software son los de *mínimos cuadrados*, *regresión*, *Anderson-Rubin y Barlett*.

En el método de regresión las puntuaciones de los factores obtenidas pueden estar correlacionadas, aun cuando se asume que los factores son ortogonales. Tampoco la varianza de las puntuaciones de cada factor es igual a 1. Con el método de Anderson-Rubin se obtienen puntuaciones de factores que están incorrelacionadas y que tienen varianza 1. Finalmente, en el método de Barlett se aplica el método de máxima verosimilitud, haciendo el supuesto de que los factores tienen una distribución normal con media y matriz de covarianzas dadas.

Sea nuestro modelo factorial $X=LF+e$, y sea $x=(x_1, \dots, x_p)$ un valor concreto de la variable medida sobre un cierto individuo de la población p -dimensional. Se trata ahora de medir el valor correspondiente de $f=(f_1, \dots, f_k)$ relativo a los k factores comunes.

Medición de componentes principales

En el caso de las componentes principales el número de factores comunes coincide con el de variables. El modelo factorial será $X=LF$ y se pueden expresar los factores directamente como combinación lineal de las variables poniendo $F=L^{-1}X$.

Si los factores son las componentes principales, la solución es todavía más directa, ya que premultiplicando por L' en el modelo factorial se tiene $L'X=L'LF$, de donde $F=(L'L)^{-1}L'X$. Además, los vectores columna de L son vectores propios ortogonales de la matriz de correlaciones R , siendo los cuadrados de sus módulos los valores propios correspondientes. Luego $L'L=D_\lambda$ es una matriz diagonal que contiene los valores propios, y $F=(D_\lambda)^{-1}L'X$, pudiéndose expresar cada componente principal según la combinación lineal:

$$F_j = \sum_{i=1}^p \frac{l_{ij}}{\lambda_j} X_i \quad j=1\dots k$$

Medición de los factores mediante estimación por mínimos cuadrados

Cuando el número de factores comunes es inferior a p , no es posible expresarlos directamente en función de las variables, es decir, no es posible expresar $f=(f_1, \dots, f_k)$ en función de $x=(x_1, \dots, x_p)$.

Si interpretamos $x=Lf+e$ como un modelo lineal donde x y L son conocidos, f son los parámetros desconocidos y e son los errores del modelo, podemos estimar f tal que sea mínimo:

$$\sum_{i=1}^p (x_i - l_{i1}f_1 - \dots - l_{ik}f_k)^2$$

La estimación de f es: $\hat{f} = (L'L)^{-1} L'x$

Medición de los factores mediante estimación por regresión

Consideramos la regresión múltiple del factor F_i sobre las variables X_1, \dots, X_p :

$$\hat{F}_i = \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p = \hat{\beta}_i X$$

F_i verifica que $E[(F_i - \hat{F}_i)^2]$ es mínimo, y los coeficientes $\hat{\beta}$ se obtienen de la relación $\hat{\beta}_i = R^{-1}\delta_i$ siendo δ_i el vector columna con las correlaciones entre el factor F_i y las variables X . Estimando F_i mediante \hat{F}_i tendremos:

$$\hat{F}_i = \delta_i' R^{-1} X$$

y considerando los m factores comunes tendremos:

$$\hat{f} = S R^{-1} x$$

siendo $S=LT$ (las columnas de T contienen las cargas factoriales de los factores oblicuos respecto a los ortogonales) la matriz de la estructura factorial. En el caso de factores ortogonales $S=L$ y tenemos:

$$\hat{f} = L' R^{-1} x$$

Medición de los factores mediante el método de Bartlett

Bartlett considera que las variables en el modelo factorial son combinación lineal de los factores comunes, mientras que los factores únicos deben ser entendidos como desviaciones de esta combinación lineal, por lo que deben ser minimizadas.

Dados x y f , los valores de los factores únicos son:

$$u_i = (x_i - \sum_{j=1}^k l_{ij} f_j) / d_i \quad i=1, \dots, p$$

Consideramos entonces la función $G = u_1^2 + \dots + u_p^2$ y, según Bartlett, hallamos f de modo que G sea mínimo. Se tiene:

$$\frac{\partial G}{\partial f_r} = 2 \sum_{i=1}^p (x_i - \sum_{j=1}^k l_{ij} f_j) \frac{(-l_{ir})}{d_i^2} = 0 \quad r=1 \dots k$$

de donde:

$$\sum_{i=1}^p x_i \frac{l_{ir}}{d_i^2} = \sum_{i=1}^p \frac{l_{ir}}{d_i^2} \sum_{j=1}^k l_{ij} f_j = 0$$

y en notación matricial: $L'D^{-2}x = L'D^{-2}L f$, realizándose la estimación de los factores mediante:

$$\hat{f} = (L'D^{-2}L)^{-1} L'D^{-2}x$$

Medición de los factores mediante el método de Anderson y Rubin

Se trata de una modificación del método de Bartlett consistente en minimizar la función $G = u_1^2 + \dots + u_p^2$ condicionada a que los factores estimados sean ortogonales, es decir, $E(\hat{F}_i \cdot \hat{F}_j) = 0$ $i \neq j$. La solución obtenida por Anderson y Rubin es:

$$\hat{f} = B^{-1} L'D^{-2}x \quad \text{con } B^2 = L'D^{-2}R D^{-2}A$$

ANÁLISIS FACTORIAL EXPLORATORIO Y CONFIRMATORIO

Una tarea fundamental en cualquier ciencia experimental es la exploración, descripción, clasificación y análisis de los objetos y fenómenos naturales. Técnicas como el análisis de componentes principales, el análisis de correspondencias, el análisis de proximidades, la taxonomía numérica, etc., son una buena herramienta para alcanzar este objetivo.

El análisis factorial nos ha permitido analizar la dimensionalidad latente en un conjunto de n variables observables, expresada a través de unos factores comunes. Hemos dedicado este Capítulo a determinar el número de factores y su influencia en las variables, siguiendo unos criterios de estructura simple, tomando como información principal la matriz de correlaciones y sin utilizar ningún otro tipo de información. Esta es la forma de análisis que ha predominado hasta los años sesenta, bajo la influencia de Thurstone y que se conoce con el nombre de **Análisis factorial exploratorio**, análisis que ha cumplido y sigue cumpliendo una meritaria labor en Psicología y otras ciencias.

La experiencia demuestra, no obstante, que la utilización a ciegas del análisis factorial exploratorio no siempre proporciona factores fácilmente interpretables. El análisis factorial realizado con un conocimiento previo de las características de los factores suele dar mejores resultados. Más que de una exploración se trata ahora de confirmar unos factores más o menos conocidos, por razones de tradición científica, porque han sido hallados en otros análisis similares, etc. Esta es, en líneas generales, la filosofía del **Análisis factorial confirmatorio**. La utilización de un método en sentido confirmatorio obliga a comprobar si las variables se ajustan a un cierto modelo o hipótesis preexistente, de forma parcial o absoluta. Normalmente se utiliza cuando una rama del conocimiento científico ha llegado a un estado de mayor sofisticación y desarrollo, interesando construir nuevas experiencias controladas, generalizar teorías, encontrar aplicaciones, etc.

El análisis factorial puede ser correctamente utilizado en sentido confirmatorio por la especial flexibilidad del modelo factorial. Esta propiedad no la tienen, en general, otros métodos multivariantes (análisis de correspondencias, análisis canónico, etc.), en los que se trata de reducir la dimensión de los datos con pérdida mínima de información. El **Análisis factorial confirmatorio** normalmente trabaja sobre factores oblicuos. Dada una matriz de correlaciones, en análisis factorial confirmatorio se parte de una supuesta estructura factorial responsable de las relaciones entre las variables. El caso más simple consiste en establecer una hipótesis sobre el número de factores comunes. En general, el tipo de hipótesis hace referencia a la naturaleza de los factores (ortogonales, oblicuos, mixtos), al número de factores comunes, o a las cargas factoriales fijas y libres del modelo factorial. Generalmente se realiza la estimación del supuesto modelo factorial confirmatorio sujeto a determinadas restricciones mediante el método de máxima verosimilitud y posteriormente se confirman las restricciones mediante un adecuado contraste de hipótesis generalmente basado en la razón de verosimilitudes. El método de máxima verosimilitud estudiado en este Capítulo para la estimación del modelo factorial y los contrastes del modelo están incluidos en las técnicas de análisis factorial confirmatorio.

COMPONENTES PRINCIPALES Y ANÁLISIS FACTORIAL CON SPSS

COMPONENTES PRINCIPALES Y ANÁLISIS FACTORIAL

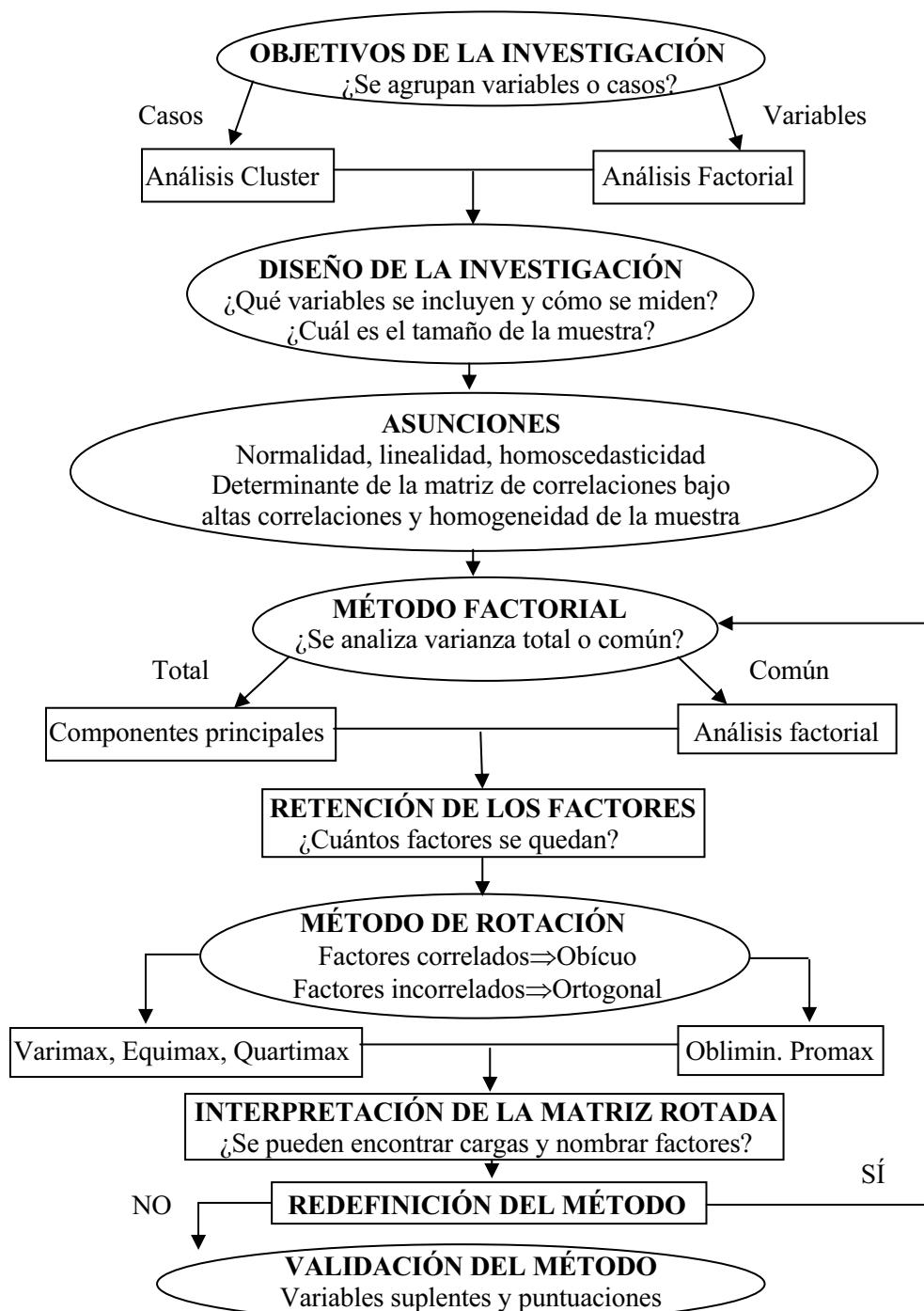
En análisis factorial y componentes principales las variables tienen que ser cuantitativas y los factores o componentes deben de ser suficientes para resumir la mayor parte de la información contenida en las variables originales.

La diferencia entre análisis en componentes principales y análisis factorial radica en que en el análisis factorial trata de encontrar variables sintéticas latentes, inobservables y aún no medidas cuya existencia se sospecha en las variables originales y que permanecen a la espera de ser halladas, mientras que en el análisis en componentes principales se obtienen variables sintéticas combinación de las originales y cuyo cálculo es posible basándose en aspectos matemáticos independientes de su interpretabilidad práctica.

En el análisis en componentes principales la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (sus componentes). Pero esto no ocurre en el análisis factorial.

En el análisis factorial sólo una parte de la varianza de cada variable original se explica completamente por las variables cuya combinación lineal la determinan (*factores comunes* F_1, F_2, \dots, F_p). Esta parte de la variabilidad de cada variable original explicada por los factores comunes se denomina *comunalidad*, mientras que la parte de varianza no explicada por los factores comunes se denomina *unicidad* (*comunalidad* + *unicidad* = 1) y representa la parte de variabilidad propia f_i de cada variable x_i . Cuando la communalidad es unitaria (unicidad nula) el análisis en componentes principales coincide con el factorial. Es decir, el análisis en componentes principales es un caso particular del análisis factorial en el que los factores comunes explican el 100% de la varianza total.

ESQUEMA GENERAL DEL ANÁLISIS FACTORIAL



SPSS Y EL ANÁLISIS FACTORIAL

Ya hemos visto que el análisis factorial intenta identificar variables subyacentes, o factores que expliquen la configuración de las correlaciones dentro de un conjunto de variables observadas. El análisis factorial se suele utilizar en la reducción de los datos para identificar un pequeño número de factores que explique la mayoría de la varianza observada en un número mayor de variables manifestas. También puede utilizarse para generar hipótesis relacionadas con los mecanismos causales o para inspeccionar las variables para análisis subsiguientes (por ejemplo, para identificar la colinealidad antes de realizar un análisis de regresión lineal).

El procedimiento de análisis factorial de SPSS ofrece un alto grado de flexibilidad ya que existen siete métodos de extracción factorial disponibles, cinco métodos de rotación de los factores incluidos el oblimin directo y el promax para rotaciones no ortogonales y tres métodos disponibles para calcular las puntuaciones factoriales, que además pueden guardarse como variables para análisis adicionales.

En cuanto a los Estadísticos, para cada variable se dispone del número de casos válidos, media y desviación típica. Para cada análisis factorial se dispone matriz de correlaciones de variables, incluidos niveles de significación, determinante, inversa; matriz de correlaciones reproducida, que incluye antiimagen; solución inicial (comunalidades, autovalores y porcentaje de varianza explicada); KMO (medida de la adecuación muestral de Kaiser-Meyer-Olkin) y prueba de esfericidad de Bartlett; solución sin rotar, que incluye saturaciones factoriales, comunalidades y autovalores; solución rotada, que incluye la matriz de configuración rotada y la matriz de transformación; para rotaciones oblicuas: para las rotaciones oblicuas: las matrices de estructura y de configuración rotadas; matriz de coeficientes para el cálculo de las puntuaciones factoriales y matriz de covarianza entre los factores. En cuanto a Diagramas se dispone del gráfico de sedimentación y del gráfico de las saturaciones de los dos o tres primeros factores.

Para realizar un análisis factorial, elija en los menús *Analizar → Reducción de datos → Análisis factorial* (Figura 6-1) y seleccione las variables y las especificaciones para el análisis (Figura 6-2). Previamente es necesario cargar en memoria el fichero de nombre MUNDO mediante *Archivo → Abrir → Datos*. Este fichero contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo. Las variables a considerar son: el índice de alfabetización (*alfabet*), el incremento de la población (*inc_pob*), la esperanza de vida femenina (*espvidaf*), la mortalidad infantil (*mortinf*), el número promedio de hijos por mujer (*fertilid*), la tasa de natalidad (*tasa_nat*), el logaritmo del PIB (*log_pib*), la población urbana (*urbana*) y la tasa de mortalidad (*tasa_mor*). Ya estamos en disposición de determinar los factores subyacentes a este conjunto de variables.

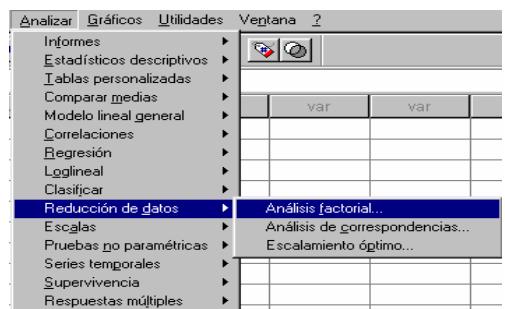


Figura 6-1



Figura 6-2

En cuanto a los datos, las variables deberán ser cuantitativas a nivel de intervalo o de razón. Los datos categóricos (como la religión o el país de origen) no son adecuados para el análisis factorial. Los datos para los cuales razonablemente se pueden calcular los coeficientes de correlación de Pearson, deberían ser adecuados para el análisis factorial. Los datos han de tener una distribución normal bivariada para cada pareja de variables, y las observaciones deben ser independientes. El modelo de análisis factorial especifica que las variables vienen determinadas por los factores comunes (los factores estimados por el modelo) y por factores únicos (los cuales no se superponen entre las distintas variables observadas); las estimaciones calculadas se basan en el supuesto de que ningún factor único está correlacionado con los demás, ni con los factores comunes.

El botón *Descriptivos* de la Figura 6-2 nos lleva a la pantalla de la Figura 6-3 en la que se establecen los estadísticos más relevantes relativos a las variables que ofrecerá el análisis. Los *Descriptivos univariados* incluyen la media, la desviación típica y el número de casos válidos para cada variable. La *Solución inicial* muestra las comunidades iniciales, los autovalores y el porcentaje de varianza explicada. En el cuadro *Matriz de correlaciones* las opciones disponibles son: *Coeficientes*, *Niveles de significación*, *Determinante*, *Inversa*, *Reproducida*, *Anti-imagen* y *KMO* y *Prueba de esfericidad de Bartlett*.

El botón *Extracción* de la Figura 6-2 nos lleva a la pantalla de la Figura 6-4 cuyo cuadro *Método* permite especificar el método de extracción factorial. Los métodos disponibles son: *Componentes principales*, *Mínimos cuadrados no ponderados*, *Mínimos cuadrados generalizados*, *Máxima verosimilitud*, *factorización de Ejes principales*, *factorización Alfa* y *factorización Imagen*. El cuadro *Analizar* permite especificar una matriz de correlaciones o una matriz de covarianza. En el cuadro *Extraer* se pueden retener todos los factores cuyos autovalores excedan un valor especificado o retener un número específico de factores. El cuadro *Mostrar* permite solicitar la solución factorial sin rotar y el gráfico de sedimentación de los autovalores. El botón *Nº máximo de iteraciones para convergencia* permite especificar el número máximo de pasos que el algoritmo puede seguir para estimar la solución.

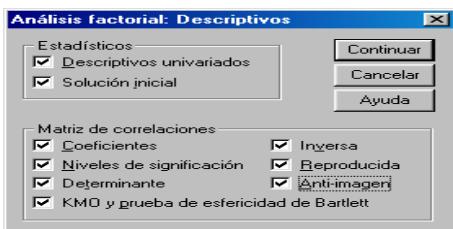


Figura 6-3

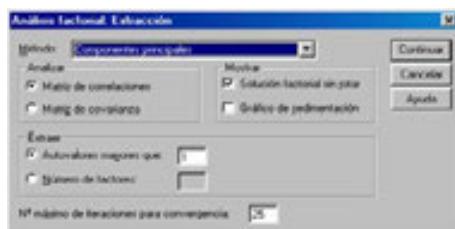


Figura 6-4

El botón *Rotación* de la Figura 6-2 lleva a la pantalla de la Figura 6-5, cuyo cuadro *Método* permite seleccionar el método de rotación factorial. Los métodos disponibles son: *Varimax*, *Equamax*, *Quartimax*, *Oblimin directo* y *Promax*. El cuadro *Mostrar* permite incluir los resultados de la solución rotada, así como los gráficos de las saturaciones para los dos o tres primeros factores. El botón *Nº máximo de iteraciones para convergencia* permite especificar el número máximo de pasos que el algoritmo elegido puede seguir para llevar a cabo la rotación.



Figura 6-5

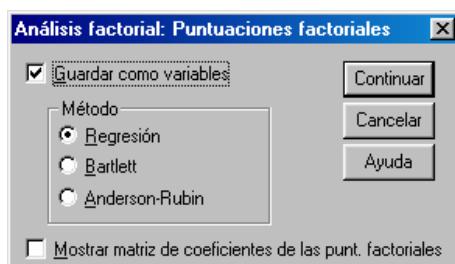


Figura 6-6

El botón *Puntuaciones* de la Figura 6-2 nos lleva a la pantalla de la Figura 6-6, cuyo botón *Guardar como variables* crea una nueva variable para cada factor en la solución final que contiene las puntuaciones. En el cuadro *Método* seleccione uno de los siguientes métodos alternativos para calcular las puntuaciones factoriales: Regresión, Bartlett o Anderson-Rubin. El botón *Mostrar matriz de coeficientes de las puntuaciones factoriales* ofrece los coeficientes por los cuales se multiplican las variables para obtener puntuaciones factoriales. También muestra las correlaciones entre las puntuaciones factoriales.

El botón *Opciones* de la Figura 6-2 nos lleva a la pantalla de la Figura 6-7, cuyo cuadro *Valores perdidos* permite especificar el tratamiento que reciben los valores perdidos. Las alternativas disponibles son: *Excluir casos según lista*, *Excluir casos según pareja* y *Reemplazar por la media*. El cuadro *Formato de visualización de los coeficientes* permite controlar aspectos de las matrices de resultados. Los coeficientes se ordenan por tamaño y se suprime aquéllos cuyos valores absolutos sean menores que el valor especificado.

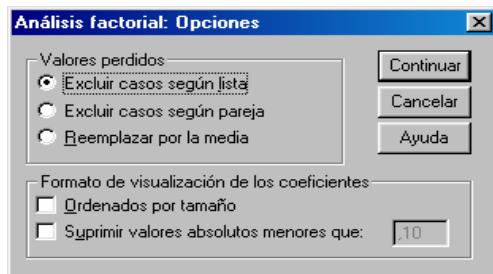


Figura 6-7

En todas las Figuras el botón *Continuar* permite pasar a la Figura 6-2 para seguir fijando especificaciones. Una vez elegidas las mismas, se pulsa el botón *Aceptar* en la Figura 6-2 para obtener los resultados del análisis factorial según se muestra en la Figura 6-8.

FACTOR /VARIABLES alfabet espidaf fertillid inc_pob log_pib mortinf tasa_mor tasa_nat urbana /MISSING LISTWISE /ANALYSIS alfabet espidaf fertillid inc_pob log_pib mortinf tasa_mor tasa_nat urbana /PRINT UNIVARIATE INITIAL CORRELATION SIG DET KMO INV REPR AIC EXTRACTION FSCORE /PLOT EIGEN /CRITERIA MINEIGEN(1) ITERATE (25) /EXTRACTION PC /ROTATION NOROTATE /SAVE REG(ALL) /METHOD=CORRELATION .			
Estadísticos descriptivos			
	Media	Desviación típica	N del análisis
Alfabetización (%)	78,14	23,06	105
Esperanza de vida femenina	69,94	10,69	105
Número promedio de hijos	3,551	1,891	105
Aumento de la población (% anual)	1,696	1,193	105
Log(10) de PIB_CAP	3,4086	,6273	105
Mortalidad infantil (muertes por 1000 nacimientos vivos)	43,317	38,370	105
Tasa de mortalidad (por 1.000 habitantes)	9,62	4,28	105
Tasa de natalidad (por 1.000 habitantes)	26,124	12,358	105
Habitantes en ciudades (%)	57,02	24,01	105

Figura 6-8

En la parte izquierda de la Figura 6-8 podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. A continuación se presenta el estadístico KMO y la prueba de Bartlett (Figura 6-9).

KMO y prueba de Bartlett		
Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,865
Prueta de esfericidad de Bartlett	Chi-cuadrado aproximado gl Sig.	1544,363 36 ,000

Figura 6-9

Se observa que el estadístico KMO vale 0,865, valor muy cercano a la unidad, lo que indica una adecuación excelente de nuestros datos a un modelo del análisis factorial. El contraste de Bartlett nos dice que no es significativa la hipótesis nula de variables iniciales incorrelacionadas, por lo tanto tiene sentido aplicar el análisis factorial.

A continuación se presentan las communalidades (Figura 6-10), el gráfico de sedimentación (Figura 6-11) y la varianza total explicada (Figura 6-12). Las communalidades iniciales valen 1 porque se ha elegido el método de componentes principales. El gráfico de sedimentación nos indica que sólo son mayores que 1 los autovalores de las dos primeras variables, con lo que estas dos variables resumirán al resto representándolas de forma coherente (serán las 2 componentes principales que resumen toda la información). La varianza total explicada muestra que las dos primeras componentes resumen el 88,175% de la variabilidad total.

Comunalidades		
	Inicial	Extracción
Alfabetización (%)	1,000	,859
Aumento de la población (% anual)	1,000	,958
Esperanza de vida femenina	1,000	,964
Mortalidad infantil (muertes por 1000 nacimientos vivos)	1,000	,941
Número promedio de hijos	1,000	,921
Tasa de natalidad (por 1.000 habitantes)	1,000	,966
Log(10) de PIB_CAP	1,000	,754
Habitantes en ciudades (%)	1,000	,706
Tasa de mortalidad (por 1.000 habitantes)	1,000	,865

Método de extracción: Análisis de Componentes principales.

Figura 6-10

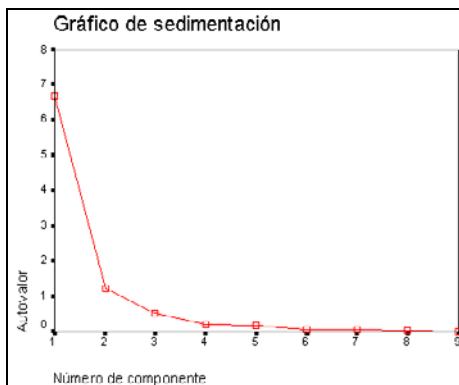


Figura 6-11

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	6,691	74,344	74,344	6,691	74,344	74,344
2	1,245	13,831	88,175	1,245	13,831	88,175
3	.532	5,907	94,083			
4	.198	2,196	96,279			
5	.174	1,932	98,211			
6	6,560E-02	.729	98,940			
7	5,587E-02	.621	99,561			
8	2,524E-02	.280	99,841			
9	1,431E-02	.159	100,000			

Método de extracción: Análisis de Componentes principales.

Figura 6-12

En la Figura 6-13 se presenta la matriz de las componentes, que es la matriz factorial que recoge la carga o ponderación de cada factor en cada una de las variables.

	Componente	
	1	2
Alfabetización (%)	-,925	-,69E-02
Aumento de la población (% anual)	,727	,656
Esperanza de vida femenina	-,963	,194
Mortalidad infantil (muertes por 1000 nacimientos vivos)	,961	-,132
Número promedio de hijos	,931	,235
Tasa de natalidad (por 1.000 habitantes)	,949	,257
Log(10) de PIB_CAP	-,867	3,876E-02
Habitantes en ciudades (%)	-,785	,299
Tasa de mortalidad (por 1.000 habitantes)	,568	-,737

Método de extracción: Análisis de componentes principales.
a. 2 componentes extraídos

Figura 6-13

Según la información de la matriz de componentes tenemos ya las variables iniciales definidas en función de las componentes (factores) de la siguiente forma:

$$\text{Alfabet} = -0,925 \text{ C1} - 0,0669 \text{ C2}$$

$$\text{Inc_pob} = 0,727 \text{ C1} + 0,6560 \text{ C2}$$

.

.

.

.

$$\text{Tasa_mor} = 0,568 \text{ C1} - 0,7370 \text{ C2}$$

Para interpretar mejor las componentes se pueden representar las variables originales en el espacio de las dos primeras componentes. Basta hacer clic en el botón *Rotación* de la Figura 6-2 y elegir *Gráfico de saturaciones* en la Figura 6-14 (y *Método Ninguno*). Se obtiene el gráfico de componentes de la Figura 6-15.

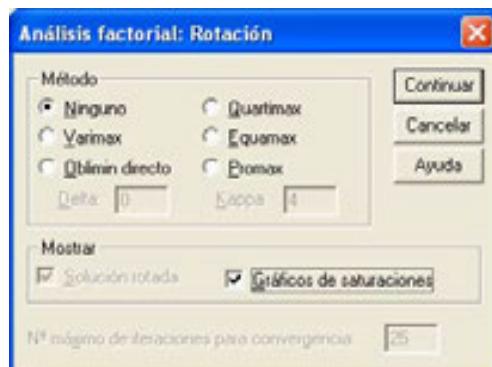


Figura 6-14

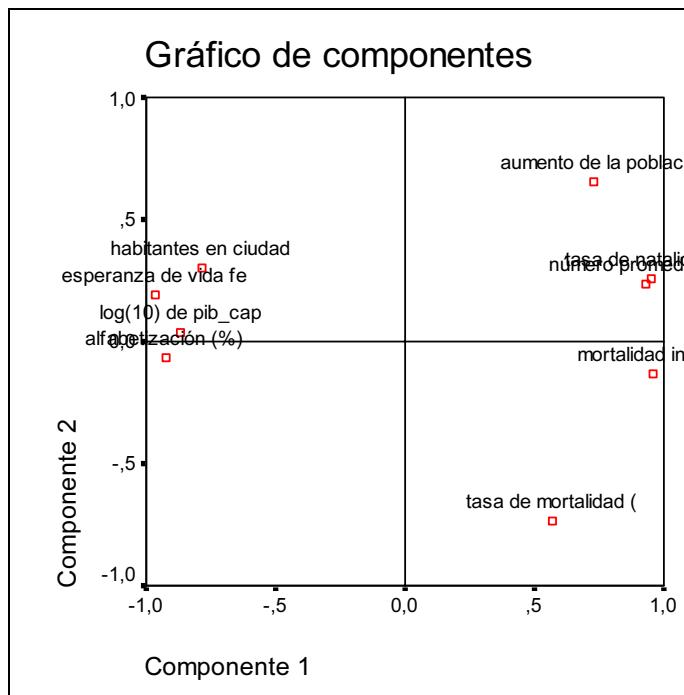


Figura 6-15

Se observa que la primera componente está correlacionada positivamente con *tasa_nat*, *mortinf* y *fertilid* (también lo estaría con *inc_pob* y *tasa_mor*, pero estos puntos están más cercanos del eje de ordenadas relativo a la segunda componente que del eje de abscisas relativo a la primera componente). La correlación es positiva porque estos puntos se sitúan a la derecha del eje de ordenadas, y la correlación es con la primera componente porque los puntos están muy cercanos al eje de abscisas. La primera componente está correlacionada negativamente con *espvidaf*, *urbana*, *log_pib* y *alfabet*. La correlación es negativa porque estos puntos se sitúan a la izquierda del eje de ordenadas, y la correlación es con la primera componente porque los puntos están muy cercanos al eje de abscisas.

La segunda componente está correlacionada positivamente con *inc_pob* y negativamente con *tasa_mor*. En este caso, la identificación de las componentes es menos clara porque las variables no están lo suficientemente cercanas al eje de ordenadas como para asegurar que se relacionen claramente con la segunda componente. Procedería entonces realizar una rotación.

Por lo tanto, inicialmente asociaríamos las variables *tasa_nat*, *mortinf*, *fertilid*, *espvidaf*, *urbana*, *log_pib* y *alfabet* con la primera componente y las variables *inc_pob* y *tasa_mor* con la segunda.

En la Figura 6-16 se presenta la matriz de coeficientes de las puntuaciones en las componentes. De sus valores podemos deducir la siguiente relación entre componentes y variables:

$$\begin{aligned} C1 &= -1,380 \text{ Alfabet} + 0,109 \text{ Inc_pob} + \dots + 0,850 \text{ Tasa_mor} \\ C2 &= -0,154 \text{ Alfabet} + 0,527 \text{ Inc_pob} + \dots - 0,592 \text{ Tasa_mor} \end{aligned}$$

	Componente	
	1	2
Alfabetización (%)	-,138	-,054
Aumento de la población (% anual)	,109	,527
Esperanza de vida femenina	-,144	,156
Mortalidad infantil (muertes por 1000 nacimientos vivos)	,144	-,106
Tasa de natalidad (por 1.000 habitantes)	,142	,206
Log(10) de PIB_CAP	-,130	,031
Número promedio de hijos	,139	,189
Habitantes en ciudades (%)	-,117	,240
Tasa de mortalidad (por 1.000 habitantes)	,085	-,502

Método de extracción: Análisis de componentes principales.

Puntuaciones de componentes.

Figura 6-16

Si echamos un vistazo a las puntuaciones que se han guardado como variables en el editor de datos con nombres *fac1_1* y *fac2_1*, vemos sus valores. Estos valores son los que pueden utilizarse en análisis posteriores (regresión, cluster, etc.) como variables sustitutas de las iniciales que las resumen en virtud del análisis factorial que acabamos de hacer.

PAÍS	fac1_1	fac2_1
Acerbaján	-,38757	,12777
Afganistán	2,42865	-1,49729
Alemania	-1,11489	-,73495
Arabia Saudí	,376031	,90962
Argentina	-,57647	,11287
Armenia	-,52297	,47847
Australia	-1,01687	,24959
Austria	-,98734	-1,05573
Bahrein	-,32838	1,57382
Bangladesh	1,42752	-,40323
Bielorusia	-,79144	-,97823
Bolivia	,55951	,46447
.	.	.
.	.	.

Si hacemos una rotación Varimax, eligiendo esta opción en *Método* de la Figura 6-14, se obtiene una solución muy parecida a la anterior sin rotar y que no aclara del todo las cosas. Ahora *tasa_mor* se asocia claramente con la segunda componente (está muy cerca del eje de ordenadas), pero *inc_pob* se asocia con la primera componente (Figura 6-17).

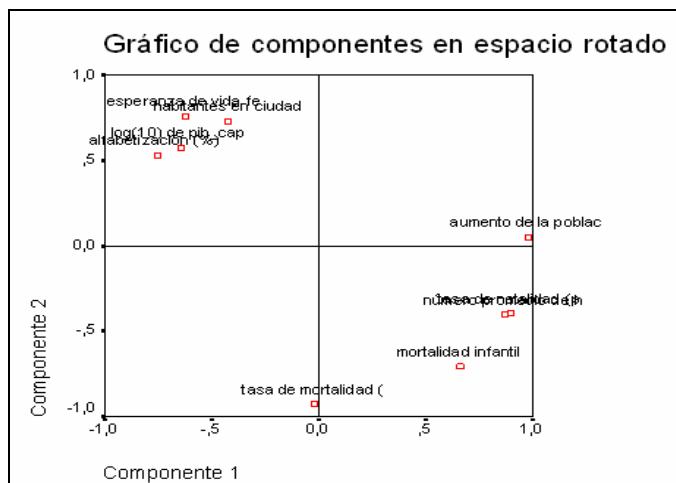


Figura 6-17

Ejercicio 6-1. Consideramos el fichero EMPRESAS.SAV que contiene información sobre empresas por países sectores de actividad. Se trata de realizar un análisis en componentes principales de todas las variables del fichero con la finalidad de reducirlas a un conjunto menor de variables con la menor pérdida de información posible.

Comenzamos eligiendo en los menús *Analizar* → *Reducción de datos* → *Análisis factorial* (Figura 6-1) y seleccionando las variables y las especificaciones para el análisis (Figura 6-18). Se incluyen todas las variables en el análisis. Previamente es necesario cargar en memoria el fichero de nombre EMPRESAS.SAV mediante *Archivo* → *Abrir* → *Datos*.

Las pantallas de los botones *Descriptivos*, *Extracción*, *Rotación*, *Puntuaciones* y *Opciones* se rellenan como se indica en las Figuras 6-19 a 6-23. Al pulsar *Continuar* y *Aceptar* se obtiene la salida del procedimiento.

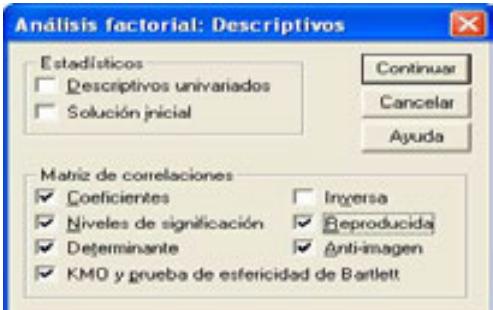
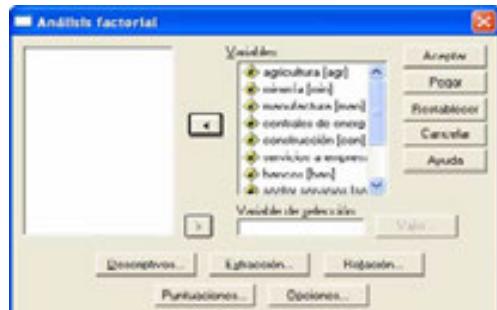


Figura 6-18

Figura 6-19

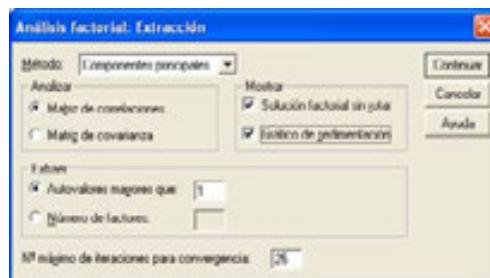


Figura 6-20

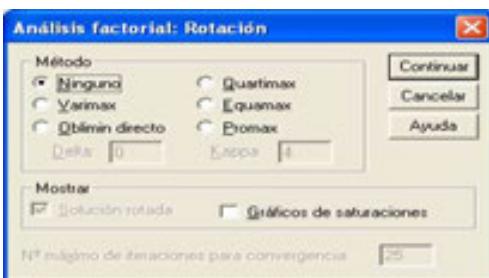


Figura 6-21

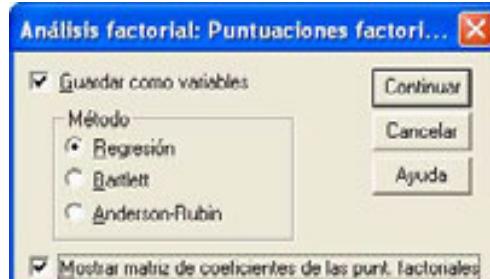


Figura 6-22

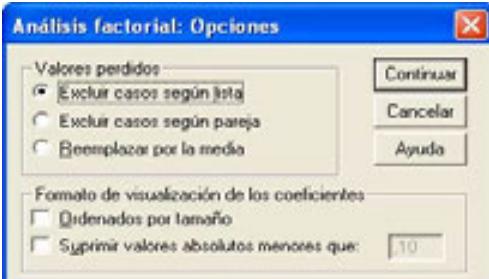


Figura 6-23

El primer elemento que se observa en la salida del procedimiento es la *matriz de correlaciones* cuyo *determinante* es $2,382 \cdot 10^{-6}$, que al ser muy pequeño indica que el grado de intercorrelación entre las variables es muy alto, condición inicial que debía cumplir el análisis en componentes principales.

Matriz de correlaciones*										
	agricultura	minería	manufactura	utilities de energía	construcción	servicios a empresas	bancos	sector servicios	transporte y comunicaciones	
Correlación	agricultura	.000	.036	-.871	-.400	-.538	-.737	-.220	-.747	.585
	minería	.036	1,000	.445	.405	.026	.397	-.443	.281	.157
	manufactura	-.871	.445	1,000	.385	.494	.704	-.356	.354	.351
	centrales de energía	.405	.405	.385	1,000	.060	.202	.110	.322	.376
	construcción	-.538	-.026	.494	.060	1,000	.356	.016	.158	.388
	servicios a empresas	-.737	.397	.204	.292	.356	1,000	.366	.572	.198
	bancos	-.220	.443	-.156	.110	.016	.366	1,000	.108	.246
	sector servicios	-.747	-.201	.154	.132	.158	.572	.108	1,000	.548
	transporte y comunicaciones	.585	.157	.251	.375	.200	.100	-.246	.560	1,000
Sig. (unilateral)	agricultura	.031	.088	.071	.002	.088	.140	.088	.088	.088
	minería	.431	.011	.020	.451	.022	.012	.082	.222	
	manufactura	.000	.011	.026	.005	.199	.214	.276	.040	
	centrales de energía	.021	.020	.026	.005	.181	.297	.260	.029	
	construcción	.002	.451	.005	.386	.037	.469	.220	.025	
	servicios a empresas	.000	.922	.159	.161	.032	.033	.001	.179	
	bancos	.140	.012	.224	.297	.489	.033	.360	.113	
	sector servicios	.000	.087	.278	.280	.279	.001	.080	.001	
	transporte y comunicaciones	.001	.222	.040	.029	.025	.179	.513	.001	

* Determinante = $2,382 \cdot 10^{-6}$

El segundo elemento que se observa en la salida del procedimiento es el *test de esfericidad de Bartlett* que permite contrastar formalmente la existencia de correlación entre las variables. Como su p-valor es 0,000, se puede concluir que existe correlación significativa entre las variables.

También se observa el *estadístico KMO*, cuyo valor tan pequeño (alejado de la unidad) indica una mala adecuación de la muestra a este análisis.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,134
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	274,053
	gl	36
	Sig.	,000

El siguiente elemento a analizar es la *matriz de correlaciones anti-imagen* formada por los coeficientes de correlación parcial entre cada par de variables cambiada de signo. Estos coeficientes deben ser bajos para que las variables comparten factores comunes. Los elementos de la diagonal de esta matriz son similares al estadístico KMO para cada par de variables e interesa que estén cercanos a la unidad. Observando la matriz vemos que no obtenemos buenos resultados.

Estadísticas anti-imagen									
	agricultura	minería	manufactura	centrales de energía	construcción	servicios a empresas	finanzas	sector servicios	transporte y comunicaciones
Covarianza anti-imagen									
agricultura	7,016<.05	,001	,000	,002	,001	,000	,000	,000	,001
minería	,001	,018	,002	,005	,011	,004	,006	,003	,012
manufactura	,000	,002	,000	,005	,001	,001	,001	,000	,002
centrales de energía	,002	,015	,005	,008	,023	,008	,013	,005	,014
construcción	,001	,011	,001	,023	,007	,002	,004	,003	,007
servicios a empresas	,000	,004	,001	,000	,002	,001	,001	,001	,003
bancos	,000	,006	,001	,013	,004	,001	,002	,001	,004
sector servicios	,000	,003	,000	,009	,002	,001	,001	,000	,002
transporte y comunicaciones	,001	,012	,002	,024	,007	,003	,004	,002	,009
Correlación anti-imagen									
agricultura	,256*	,876	,000	,002	,003	,000	,000	,000	,000
minería	,875	,564*	,875	,006	,871	,877	,878	,875	,463
manufactura	,000	,873	,146*	,000	,861	,868	,866	,866	,867
centrales de energía	,002	,026	,000	,100*	,003	,004	,079	,095	,047
construcción	,003	,015	,001	,003	,000*	,000	,000	,004	,011
servicios a empresas	,000	,077	,000	,004	,000	,154*	,007	,000	,000
bancos	,000	,078	,000	,019	,000	,007	,069*	,007	,000
sector servicios	,000	,075	,000	,005	,004	,000	,007	,151*	,003
transporte y comunicaciones	,007	,063	,007	,007	,071	,000	,000	,003	,136*

* Medida de adecuación muestral

Ahora analizaremos el número de componentes con el que nos quedaremos, que normalmente son las relativas a valores propios mayores que la unidad. Observando la tabla de la varianza total explicada vemos que la primera componente explica un 38,7% de la varianza total y las dos siguientes un 23,6% y un 12,2% (un 74,6% entre las tres). El *gráfico de sedimentación* (Figura 6-24) muestra que sólo hay tres componentes con autovalor mayor que 1.

Varianza total explicada

Componente	Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado
1	3,487	38,746	38,746
2	2,130	23,669	62,415
3	1,099	12,211	74,625

Método de extracción: Análisis de Componentes principales.

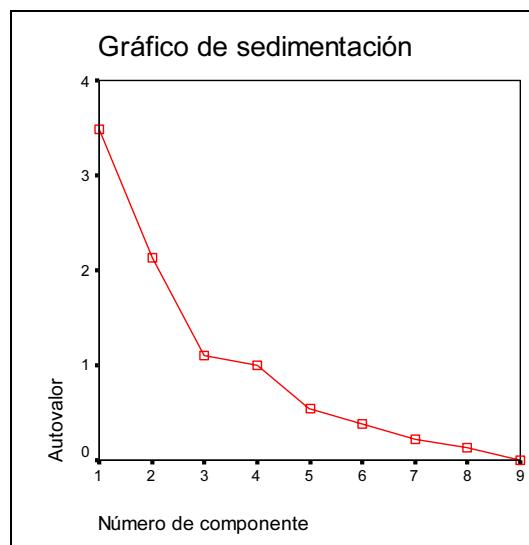


Figura 6-24

Ahora nos preocupamos de un tema tan importante como es el utilizar las variables que se agrupan en torno a cada componente y cuya combinación lineal define la componente. Para ello representamos cada una de las nueve variables por medio de los tres factores extraídos utilizando la matriz de componentes. Podemos escribir lo siguiente:

$$\begin{aligned}
 \text{Agricultura} &= -0,978F1 + 0,078F2 - 0,510F3 \\
 \text{Minería} &= -0,020F1 + 0,902F2 + 0,211F3 \\
 &\cdot \\
 &\cdot \\
 \text{T. y comunicaciones} &= 0,685F1 + 0,296F2 - 0,393F3
 \end{aligned}$$

Para ver qué variables se agrupan en cada componente (factor) hay que observar las variables cuyas cargas sean altas en un factor y bajas en los otros (valores menores que 0,25 suelen considerarse bajos).

En la primera componente está representada claramente Agricultura y en la segunda Minería (sus valores en la matriz de componentes son muy altos). Sector servicios y Transporte y comunicaciones están representadas en las tres componentes, lo mismo que Centrales de energía. Servicios a empresas y Manufacturas están representadas en la primera y segunda componentes y Bancos en la segunda y la tercera. Se observa entonces que es difícil agrupar las variables en componentes con lo que procedería realizar una rotación, que se realizará posteriormente.

Matriz de componentes^a

	Componente		
	1	2	3
agricultura	-,978	,078	-,051
minería	-,002	,902	,211
manufactura	,649	,518	,158
centrales de energía	,478	,381	,588
construcción	,607	,075	-,161
servicios a empresas	,708	-,511	,121
bancos	,139	-,662	,616
sector servicios	,723	-,323	-,327
transporte y comunicaciones	,685	,296	-,393

Método de extracción: Análisis de componentes principales.

a. 3 componentes extraídos

Es importante observar que la suma de los cuadrados de los elementos de las columnas de la matriz de componentes es igual a los valores propios significativos

A continuación se analiza la communalidad de cada variable (suma de los cuadrados de sus cargas factoriales definidas en la matriz de componentes) después de la extracción de los factores (componentes). La communalidad es la parte de variabilidad de cada variable explicada por los factores. Antes de la extracción de los factores la communalidad de cada variable es la unidad, e interesa que después de la extracción siga siendo alta.

Comunalidades

	Extracción
agricultura	,965
minería	,858
manufactura	,714
centrales de energía	,719
construcción	,400
servicios a empresas	,776
bancos	,837
sector servicios	,735
transporte y comunicaciones	,711

Método de extracción: Análisis de Componentes principales.

Es posible calcular los coeficientes de correlación entre cada dos variables después de que estén en función de las componentes, denominados *coeficientes de correlación reproducidos*. Estos coeficientes de correlación reproducidos no tienen porqué coincidir con los de la matriz de correlaciones inicial, pero no deben diferenciarse en más de 0,05 (residuos menores que 0,05), porque entonces la bondad del modelo factorial será discutible. A continuación se presenta la *matriz de correlaciones reproducidas* en la que se observa que un 61% de los errores son mayores que 0,05, lo que indica que la bondad del modelo es discutible.

Correlaciones reproducidas									
	agricultura	minería	manufactura	centrales de energía	construcción	servicios a empresas	bancos	sector servicios	transporte y comunicaciones
Correlación reproducida									
agricultura	,365 ^a	,267	-,067	-,067	-,566	-,738	,719	-,716	,037
minería	,262	,399 ^b	,499	,499	,032	,427	,499	,262	,382
manufactura	,362	,499	,214 ^b	,600	,497	,212	,196	,290	,326
centrales de energía	-,487	,486	,600	,719 ^b	,234	,214	,176	,050	,208
construcción	-,585	,037	,487	,794	,488 ^b	,377	,064	,488	,581
servicios a empresas	,738	,437	,213	,214	,372	,718 ^b	,511	,328	,296
bancos	,279	,499	,196	,176	,064	,511	,337 ^b	,313	,343
sector servicios	,216	,342	,260	,050	,488	,638	,113	,736 ^b	,329
transporte y comunicaciones	,327	,192	,508	,208	,501	,298	,343	,529	,311 ^b
Residual ^c									
agricultura		-,026	-,069	,067	,047	,001	-,001	,031	,067
minería		,026	-,054	-,061	-,050	,041	,025	,001	,025
manufactura		,068	-,056		,215	,067	-,010	,6,92E-05	,096
centrales de energía		,067	-,061		,215	,064	,012	,066	,187
construcción		,043	,058	,087	-,164		,016	,080	,114
servicios a empresas		,001	,041	,010	,012		,016	,006	,009
bancos		,001	,025	6,92E-05	,066	,000	,045	,006	,097
sector servicios		,031	,001	,098	,102	,009	,006	,006	,019
transporte y comunicaciones		,062	,025	,195	,167	,014	,009	,091	,039

Método de extracción: Análisis de componentes principales.
 a: Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 22 (81,0%) residuos más redundantes con valores absolutos mayores que 0,05.
 b: Componentes reproducidos
 c: Componentes residuales

Ahora calcularemos las *puntuaciones factoriales*, que no son más que los valores que toman cada uno de los individuos en las tres componentes seleccionadas. Serán entonces tres variables sustitutas de las iniciales que representan su reducción y que recogen el 74,6% de la variabilidad total. Estas variables son las que se utilizarán como sustitutas de las iniciales para análisis posteriores tales como el análisis de la regresión con problemas de multicolinealidad y el análisis cluster. SPSS incorpora estas variables al conjunto de datos si así se le pide en la Figura 6-8.

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

	Componente		
	1	2	3
agricultura	-,280	,037	-,046
minería	-,001	,423	,192
manufactura	,186	,243	,144
centrales de energía	,137	,179	,535
construcción	,174	,035	-,146
servicios a empresas	,203	-,240	,110
bancos	,040	-,311	,560
sector servicios	,207	-,152	-,298
transporte y comunicaciones	,196	,139	-,358

Método de extracción: Análisis de componentes principales.
 Puntuaciones de componentes.

De la matriz de coeficientes de las puntuaciones en las componentes anteriores podemos deducir la siguiente relación entre componentes y variables:

$$\begin{aligned} C1 &= -2,800 \text{ Agricultura} - 0,010 \text{ Minería} + \dots + 0,196 \text{ T. y comunic.} \\ C2 &= 0,037 \text{ Agricultura} + 0,423 \text{ Minería} + \dots + 0,139 \text{ T. y comunic.} \\ C3 &= -0,046 \text{ Agricultura} + 0,192 \text{ Minería} + \dots - 0,358 \text{ T. y comunic.} \end{aligned}$$

Cuando se realizó el análisis de la matriz de componentes se observó que es difícil agrupar las variables en componentes, con lo que procedería realizar una rotación. Realizaremos una *rotación Varimax* que tiene la propiedad de que los factores siguen siendo incorrelados. Para ello hacemos clic en el botón *Rotación* de la pantalla de entrada del Análisis Factorial (Figura 6-14) y rellenamos la pantalla *Análisis factorial: Rotación* como se indica en la Figura 6-25.

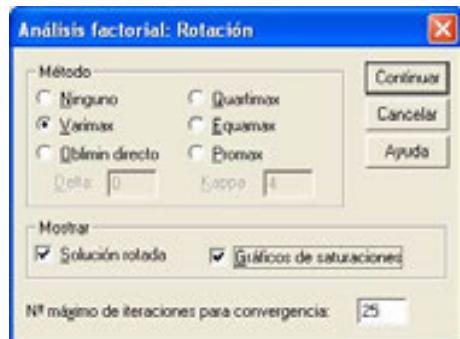


Figura 6-25

Al hacer clic en *Continuar* y *Aceptar*, se obtiene la *Matriz de componentes rotados*, que muestra cómo la variable Servicios a empresas se sitúa en la primera componente, la variable Centrales de energía se sitúa en la segunda componente y la variable Construcción se sitúa en la primera componente. A pesar de la rotación, no se ven claros los grupos de variables.

Matriz de componentes rotados^a

	Componente		
	1	2	3
agricultura	-.871	-.343	-.299
minería	-.186	.743	-.520
manufactura	.465	.692	-.136
centrales de energía	.146	.809	.207
construcción	.607	.174	-.030
servicios a empresas	.643	-.006	.602
bancos	-.060	-.005	.913
sector servicios	.824	-.157	.177
transporte y comunicaciones	.751	.205	-.325

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

El gráfico de componentes en el espacio rotado (Figura 6-26) tampoco ayuda mucho a la detección de los grupos de variables. En este gráfico, dos variables correladas positivamente forman un ángulo desde el origen de 0 grados, de 180 si lo están negativamente y de 90 si están incorreladas.

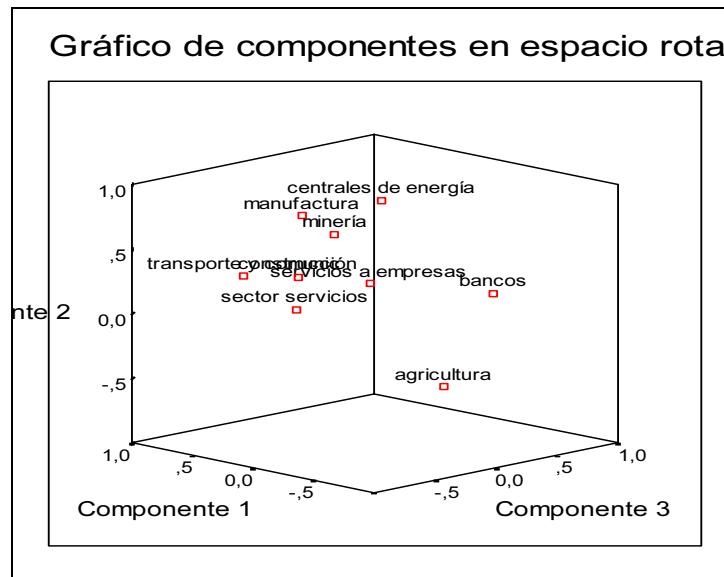


Figura 6-26

Este gráfico se puede descomponer en varios gráficos bidimensionales haciendo doble clic sobre él y elegiendo *Galería → Dispersion* (Figura 6-27). A continuación se selecciona *Simple* y se hace clic en *Reemplazar* (Figura 6-28) para obtener la pantalla de la Figura 6-29 en cuyo campo *Mostrar en el eje* se sitúan las componentes a graficar bidimensionalmente. Al hacer clic en *Aceptar*, se obtiene el gráfico de la Figura 6-30. Al hacer la rotación aparece una nueva matriz de componentes.

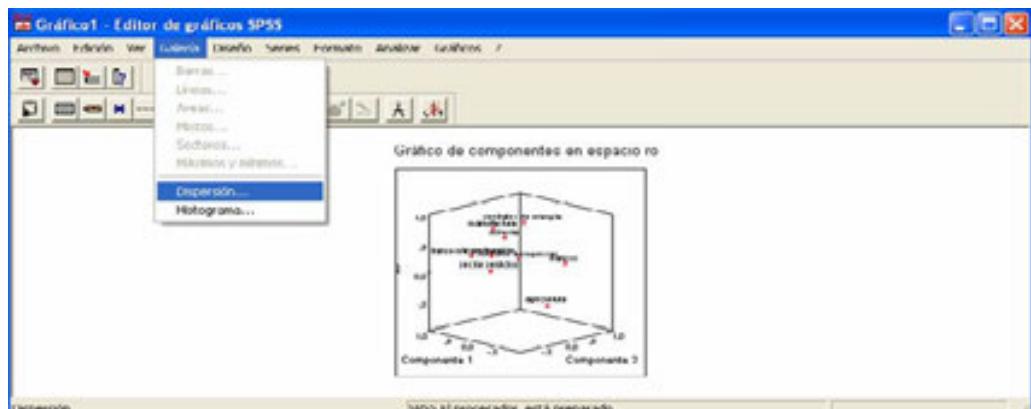


Figura 6-27

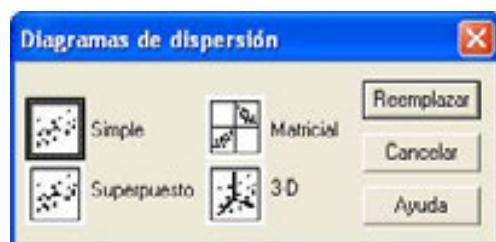


Figura 6-28

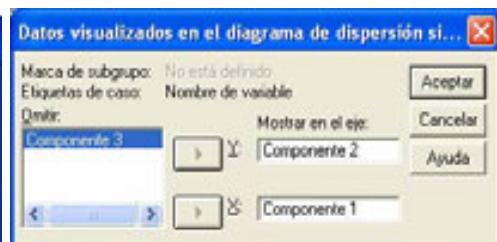


Figura 6-29

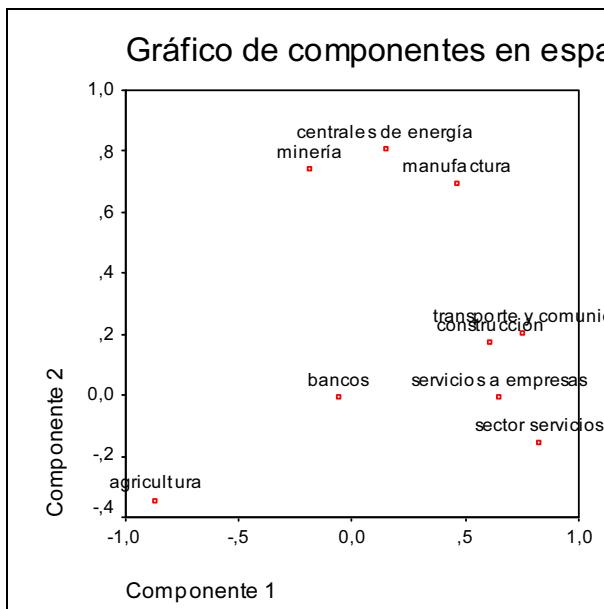


Figura 6-30

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

	Componente		
	1	2	3
agricultura	-,238	-,109	-,117
minería	-,125	,408	-,184
manufactura	,083	,325	-,044
centrales de energía	-,118	,512	,247
construcción	,214	-,004	-,084
servicios a empresas	,163	-,017	,290
bancos	-,164	,160	,600
sector servicios	,327	-,215	-,039
transporte y comunicaciones	,311	-,060	-,292

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

Puntuaciones de componentes.

Ejercicio 6-2. Una empresa especializada en el diseño de automóviles de turismo desea estudiar cuáles son los deseos del público que compra automóviles. Para ello diseña una encuesta con 10 preguntas donde se le pide a cada uno de los 20 encuestados que valore de 1 a 5 si una característica es o no muy importante. Los encuestados deberán contestar con un 5 si la característica es muy importante, un 4 si es importante, un 3 si tiene regular importancia, un 2 si es poco importante y un 1 si no es nada importante. Las 10 características (V_1 a V_{10}) a valorar son: precio, financiación, consumo, combustible, seguridad, confort, capacidad, prestaciones, modernidad y aerodinámica. El fichero 6-2.sav recoge los datos. Realizar un análisis factorial que permita extraer unos factores adecuados a los datos que resuman correctamente la información que contienen.

Comenzamos rellenando la pantalla de entrada del procedimiento *Análisis factorial* como se indica en la Figura 6-31 (se incluyen las diez variables en el análisis factorial). Las pantallas de los botones *Descriptivos*, *Extracción* y *Puntuaciones* se rellenan como se indica en las Figuras 6-32 a 6-34. Al pulsar *Continuar* y *Aceptar* se obtiene la salida de las Figuras 6-35 a 6-39.



Figura 6-31

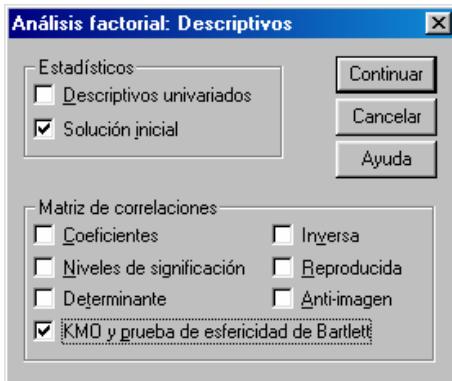


Figura 6-32

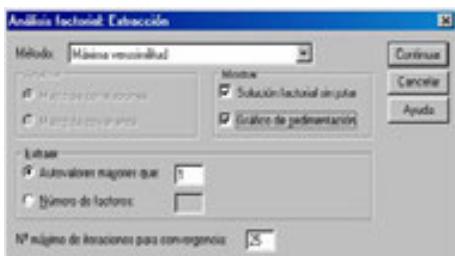


Figura 6-33

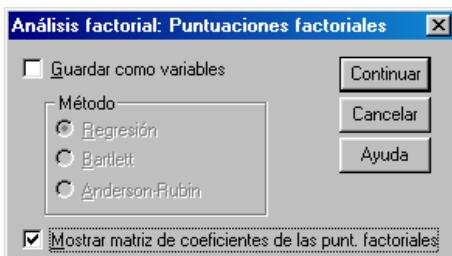


Figura 6-34

```

SAVE OUTFILE='C:\Archivos de programa\SPSS\15-1.sav'
/COMPRESSED.
FACTOR
/VARIABLES v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 /MISSING LISTWISE /ANALYSIS v1
v2 v3 v4 v5 v6 v7 v8 v9 v10
/PRINT INITIAL KMO EXTRACTION FSCORE
/PLOT EIGEN
/CRITERIA MINEIGEN(1) ITERATE (25)
/EXTRACTION ML
/ROTATION NOROTATE .

```

Análisis factorial**KMO y prueba de Bartlett**

Medida de adecuación muestral de Kaiser-Meyer-Olkin.	,700
Prueba de esfericidad de Bartlett	
Chi-cuadrado aproximado	163,466
gl	45
Sig.	,000

Figura 6-35

Comunalidades		
	Inicial	Extracción
V1	,875	,752
V2	,905	,845
V3	,781	,677
V4	,848	,873
V5	,800	,762
V6	,845	,826
V7	,534	,298
V8	,918	,717
V9	,907	,760
V10	,794	,738

Método de extracción: Máxima verosimilitud.

Figura 6-36

Varianza total explicada						
Factor	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	5,701	57,011	57,011	5,464	54,636	54,636
2	2,069	20,692	77,703	1,785	17,846	72,482
3	,720	7,205	84,908			
4	,548	5,478	90,386			
5	,316	3,158	93,544			
6	,271	2,707	96,251			
7	,146	1,464	97,715			
8	,128	1,280	98,995			
9	6,836E-02	,684	99,679			
10	3,210E-02	,321	100,000			

Método de extracción: Máxima verosimilitud.

Figura 6-37

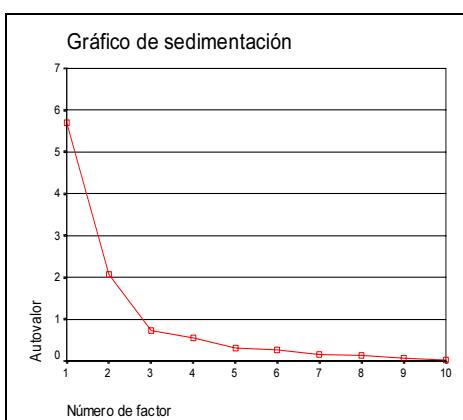


Figura 6-38

	Matriz factorial ^a	
	1	2
V1	,867	-3,47E-02
V2	,910	,131
V3	,820	-6,06E-02
V4	,933	-5,02E-02
V5	-,557	,672
V6	-,236	,878
V7	,217	,501
V8	-,829	-,172
V9	-,849	-,197
V10	-,721	-,467

Método de extracción: Máxima verosimilitud.
a. 2 factores extraídos. Requeridas 10 iter

Figura 6-39

En la Figura 6-35 se observa que el estadístico KMO vale 0,7, valor cercano a la unidad, lo que indica una adecuación correcta de nuestros datos a un modelo del análisis factorial. El p-valor del contraste de Bartlett nos dice que no es significativa la hipótesis nula de variables iniciales incorrelacionadas, por lo tanto tiene sentido aplicar el análisis factorial. A continuación se presentan las Comunalidades (Figura 6-36), la Varianza total explicada (Figura 6-37) y el Gráfico de sedimentación (Figura 6-38). El gráfico de sedimentación nos indica que sólo son mayores que dos autovalores que explican el 77,7% de la variabilidad total, con lo que habrá dos factores que resumirán a todas las variables representándolas de forma coherente.

Para identificar los factores se utiliza la matriz factorial de la Figura 6-39 que nos da los coeficientes de correlación de cada uno de los dos factores con las 10 variables. El factor 1 lo formarán las variables que tengan correlación más fuerte en módulo con dicho factor (V1, V2, V3, V4, V8, V9 y V10). El factor 2 lo formarán las variables con correlación más fuerte con dicho factor (V5, V6 y V7).

Si atendemos a la naturaleza de las variables que componen el primer factor, se observa que se trata de un factor que une la economicidad (correlaciones altas positivas) con el escaso interés en que el coche tenga aire deportivo (correlaciones altas negativas). Este factor explica por sí solo el 57% de la varianza, lo que quiere decir que discrimina muy bien al colectivo. Por otro lado, si atendemos a la naturaleza de las variables que componen el segundo factor, se observa que se trata de un factor que mide la utilidad del coche. Este factor explica por sí solo el 20,7% de la varianza (siempre menos que el primero).

En este caso, la delimitación de las variables que forman cada factor es clara, pero no siempre es así. Cuando hay dificultad de formación de los factores será necesario rotarlos para intentar aclarar la composición de los factores. Para realizar una rotación se utiliza el botón *Rotación* de la Figura 6-31, en cuya pantalla de entrada (Figura 6-40) se elige el método de rotación. Al pulsar *Continuar* y *Aceptar* se obtiene la matriz de factores rotados de la Figura 6-41, que agruparía las variables en factores igual que la matriz sin rotar. Varianzas explicadas, KMO y autovalores tampoco difieren.

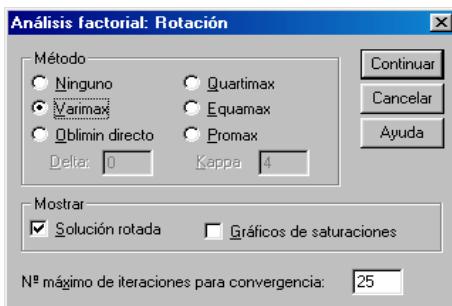


Figura 6-40

	Factor	
	1	2
V1	.849	-.175
V2	.919	-1,85E-02
V3	.800	-.193
V4	.912	-.201
V5	.440	.754
V6	-9,04E-02	.905
V7	.296	.459
V8	.848	-3,47E-02
V9	.870	-5,66E-02
V10	.787	-.343

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 3 iteraciones.

Figura 6-41

Ejercicio 6-3. Para estudiar las ventas de las empresas españolas se consideran 7 variables ratio R1 a R7. Concretamente los ratios son beneficios/recursos propios (R1), cash-flow/ventas (R2), inmovilizado/activos totales (R3), ventas/activos totales (R4), ventas/plantilla (R5), beneficios/capital social (R6) y beneficios/ventas (R7) que caracterizan a las empresas españolas con mayores ventas. Se trata de resumir estos ratios por un número menor de factores con mínima pérdida de información que tengan la suficiente calidad para seguir agrupando a las empresas según sus ventas. ¿Sería coherente identificar un factor financiero, un factor estructural y un factor de rentabilidad? El fichero 6-3.sav contiene la información de las variables.

El resumen se llevará a cabo mediante análisis factorial, abriendo el fichero 6-3.sav y rellenando la pantalla de entrada del procedimiento *Análisis factorial* (*Analizar → Reducción de datos → Análisis factorial*) como se indica en la Figura 6-42. Las pantallas de los botones *Descriptivos*, *Extracción*, *Puntuaciones factoriales* y *Rotación* se llenan como se indica en las Figuras 6-43 a 6-46. Se utiliza el método factorial *Análisis Alfa* para extraer tres factores (el tercero corresponde a un autovalor menor que 1, pero que está muy cercano a 1, y por eso se incluye). Al pulsar *Continuar* y *Aceptar* se observa una cierta indefinición entre tomar las 2 o 3 primeras componentes. El gráfico de sedimentación tiene sólo 2 valores propios mayores que uno con un tercero muy próximo (Figura 6-49) y las dos primeras componentes sólo explican el 64,8% de la varianza (Figura 6-48). El determinante de la matriz de correlaciones es muy bajo y el estadístico KMO tiene valor alto, lo que indica una buena adecuación muestral al análisis factorial (Figura 6-47). Una posición conservadora sería tomar los tres primeros factores.



Figura 6-42

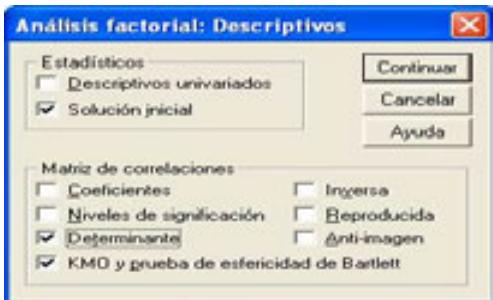


Figura 6-43

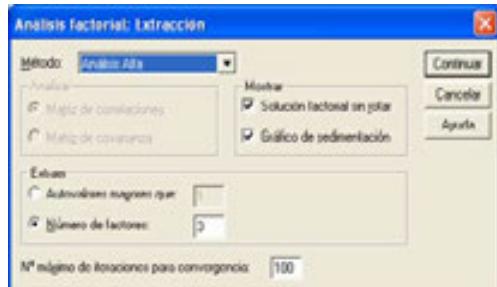


Figura 6-44

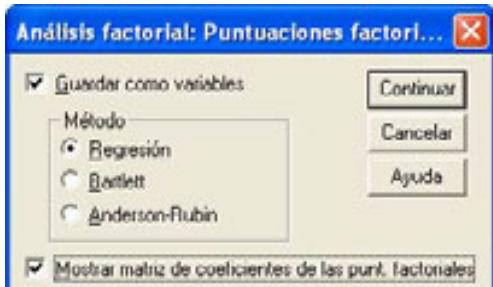


Figura 6-45

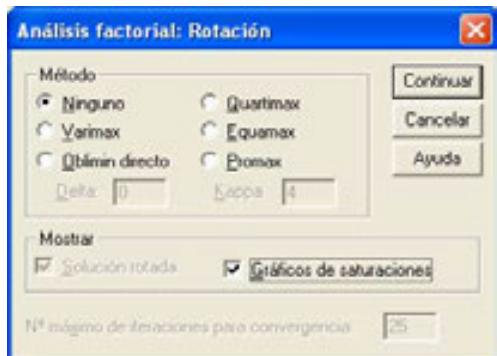


Figura 6-46

Matriz de correlaciones*	
a. Determinante = 1,265E-09	
KMO y prueba de Bartlett	
Medida de adecuación muestral de Kaiser-Meyer-Olkin	,876
Prueba de esfericidad de Bartlett	
Chi-cuadrado aproximado	4565,530
gl	21
Sig.	,000

Figura 6-47

Varianza total explicada			
Factor	Autovalores iniciales		
	Total	% de la varianza	% acumulado
1	2,946	42,092	42,092
2	1,591	22,727	64,819
3	,996	14,223	79,042
4	,892	12,740	91,782
5	,538	7,682	99,464
6	,038	,536	100,000
7	1,506E-08	2,151E-07	100,000

Método de extracción: Máxima verosimilitud.

Figura 6-48

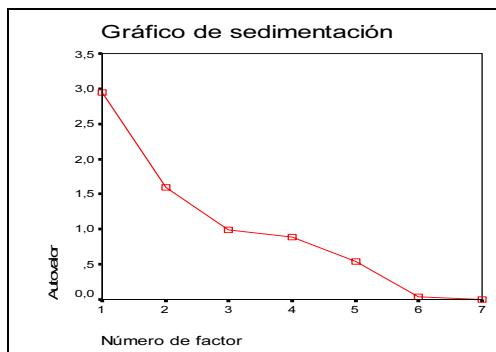


Figura 6-49

Comunalidades		
	Inicial	Extracción
R1	,004	,004
R2	,944	,964
R3	,420	,459
R4	,292	,601
R5	,093	,124
R6	1,000	,995
R7	1,000	,995

Método de extracción: Factorización Alfa.

Figura 6-50

	Matriz factorial		
	Factor 1	Factor 2	Factor 3
R1	,054	-,025	-,020
R2	,389	,901	-,012
R3	-,512	,381	,226
R4	,659	-,385	,139
R5	,280	-,056	,207
R6	,595	,797	-,075
R7	,595	,797	-,075

Método de extracción: Factorización Alfa.
a. 3 factores extraídos. Requeridas 10 iteraciones.

Figura 6-51

La Figura 6-50 muestra las comunalidades y la Figura 6-51 muestra las cargas factoriales. Cargas factoriales altas en valor absoluto de una variable sobre un factor indican que hay mucho en común entre la variable y el factor. Hay autores que sostienen que cargas mayores que 0,6 asocian a la variable con el factor, mientras que otros sostienen que es suficiente un valor superior a 0,4. En nuestro caso no hay forma clara de asociar nuestras variables a los factores, por lo que haremos una rotación. Además, el gráfico tridimensional de factores no despeja las dudas (Figura 6-52).

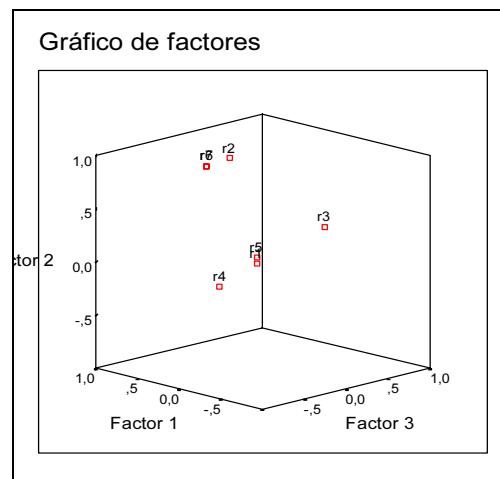


Figura 6-52



Figura 6-53

Si ahora realizamos la rotación de los factores por el método Varimax haciendo clic en el botón *Rotación* y marcando *Varimax* en la pantalla resultante (Figura 6-53), al hacer clic en *Continuar* y *Aceptar*, obtenemos las cargas factoriales de la matriz factorial rotada (Figura 6-54).

Matriz de factores rotados^a

	Factor		
	1	2	3
R1	,007	,060	,015
R2	,971	-,143	,029
R3	,052	-,661	-,136
R4	-,025	,575	,519
R5	,065	,116	,327
R6	,989	,084	,099
R7	,989	,084	,098

Método de extracción: Factorización Alfa.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 5 iteraciones.

Figura 6-54



Figura 6-55

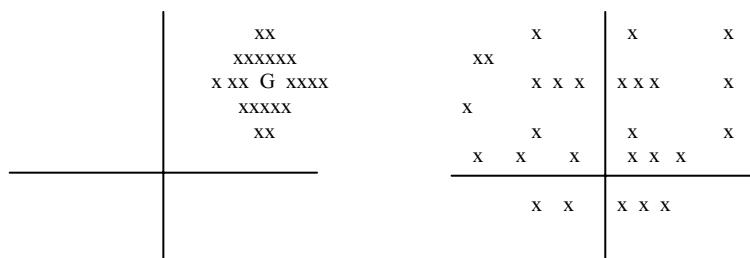
La matriz de factores rotados muestra claramente que al primer factor se asocian las variables R2, R6 y R7 con cargas factoriales mayores que 0,9. Al segundo factor se asocia R3 y al tercero se asocian R4 y R5. Como nos queda suelta R1, la asociamos al factor para el que presenta mayor carga, es decir, a R2. También podía haberse asociado R4 al segundo factor, pero esta asociación está más clara para el tercer factor, ya que es su única carga realmente alta. A estas mismas conclusiones puede llevarnos el gráfico de saturaciones en el espacio factorial rotado (Figura 6-55)

Dada la naturaleza de las variables, podemos decir que el primer factor (R2, R6 y R7) es un FACTOR FINANCIERO relativo a la distribución de los beneficios y flujo de caja, el segundo factor (R1 y R3) es un FACTOR ESTRUCTURAL relativo a recursos propios, inmovilizado y activos totales y el tercer factor (R4 y R5) es un factor de rentabilidad relativo a la distribución de las ventas.

MÉTODOS FACTORIALES EN GENERAL. ANÁLISIS DE CORRESPONDENCIAS

CANTIDAD DE INFORMACIÓN Y DISTANCIAS

El objetivo principal del análisis de datos suele ser resumir y sintetizar la información contenida en una gran tabla de datos, de manera que, permitiendo una pequeña pérdida de información, se produzca una ganancia en significación. Para poder vigilar la calidad de los resultados, así como para diseñar el método de análisis, es necesario definir lo que entendemos por *cantidad de información*. Existen diversas formas de medir la cantidad de información. Para considerarlas, comencemos por imaginarnos una tabla en la que se miden dos variables para n individuos. Gráficamente podemos representar la tabla mediante un plano cuyos ejes representan las dos variables respectivamente, y cada punto representa un individuo cuyas coordenadas son los valores que toma para cada una de las variables. De esta forma podríamos obtener representaciones de los puntos en un plano formando nubes como las presentadas en las dos Figuras siguientes:



En la primera Figura casi todos los puntos son semejantes, y toman valores próximos para ambas variables. No hay muchas diferencias entre unos individuos y otros y cada uno de ellos individualmente no aporta mucha información al colectivo. Se podrían representar todos ellos bastante bien por su centro de gravedad G, que los resume adecuadamente.

En la Figura de la derecha los individuos están más dispersos, más separados. Su centro de gravedad G los representa peor, los resume mal, y varía mucho con la introducción o supresión de un individuo. Todos son muy diferentes e individualmente aportan mucha información al colectivo. De esta forma podríamos decir intuitivamente que los puntos de la primera Figura contienen poca información, mientras que los puntos de la Figura de la derecha contienen mucha.

Existen diversas formas matemáticas de medir la cantidad de información, algunas de las cuales se utilizan en los métodos factoriales. Todas las medidas de la información incluyen una medida de la distancia entre los puntos, por lo que es necesario definir también el concepto de **distancia**. La distancia entre dos individuos o variables mide el grado de asociación o semejanza entre éstas. Existen distintas medidas de la distancia y todas ellas cumplen los siguientes axiomas:

1. $\forall i, i' \quad d_{ii'} > 0 \text{ y } d_{ii} = 0$ (la distancia nunca es negativa y la distancia de un punto a sí mismo es cero).
2. $\forall i, i' \quad d_{ii'} = d_{i'i}$ (la distancia es simétrica).
3. $\forall a \neq b \neq c \quad d_{ac} \leq d_{ab} + d_{bc}$ (desigualdad triangular).

La distancia más utilizada con variables cuantitativas es la **distancia euclídea**. Sean i e i' dos individuos en los que se han medido p variables. Estos individuos están representados por los valores que toman para el conjunto de variables x_i y $x_{i'}$. La distancia euclídea al cuadrado se mide mediante la suma de las diferencias, el cuadrado de los valores de cada variable en los dos individuos a través de la fórmula siguiente:

$$d_{ii'}^2 = \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Dos individuos que toman valores próximos para todo el conjunto de variables tendrán una distancia pequeña (son semejantes).

Otra función de distancia utilizada en algunos métodos multivariantes es la **distancia χ^2** . Se trata de una distancia entre distribuciones o perfiles que se utiliza cuando se analizan tablas de frecuencias. La distancia χ^2 entre dos filas i e i' de términos k_{ij} y $k_{i'j}$ se calcula mediante la siguiente fórmula:

$$d_{ii'}^2 = \sum_{j=1}^p \frac{1}{k_j/k} \left(\frac{k_{ij}}{k_i} - \frac{k_{i'j}}{k_{i'}} \right)^2$$

donde k_{ij} es la frecuencia de asociación de i y j , k_i es la frecuencia con que se ha presentado i , $k_i = \sum_j k_{ij}$ y $k = \sum_{ij} k_{ij}$. Por lo tanto se trata de una distancia euclídea ponderada.

Existen diversas funciones de distancia, y cada una tiene unas propiedades que la hacen más adecuada a un tipo de datos o de análisis.

Una **medida de la información** de una tabla de datos de n individuos y p variables es la suma de los cuadrados de distancias de los individuos i al origen. La fórmula es la siguiente:

$$I = \sum_{i=1}^n d^2(i, 0)$$

pero, generalmente, el origen suele hacerse coincidir con el centro de gravedad G , con lo que la información se mide mediante la fórmula:

$$I = \sum_{i=1}^n d^2(i, G)$$

y si las variables son métricas, se puede utilizar la distancia euclídea, con lo que se tiene:

$$I = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - G_j)^2$$

fórmula que puede expresarse como:

$$I = \sum_{j=1}^p \sum_{i=1}^n (x_{ij} - G_j)^2$$

es decir, como la suma de las varianzas de las variables, razón por la que esta medida de la información se denomina **varianza total**.

Este **criterio de la varianza** se utiliza en muchos métodos factoriales para medir la cantidad de información de una tabla o la cantidad de información mantenida después de un análisis.

También suele utilizarse para medir la cantidad de información la **inercia de la nube de puntos $I(N)$ con relación al centro de gravedad G** (medida de la dispersión de los puntos en torno a su centro), cuya expresión es la siguiente:

$$I_G(N) = \sum_{i=1}^n p_i d^2(i, G)$$

Esta fórmula de la inercia de la nube de puntos representa la suma de las distancias al cuadrado de los puntos al centro de gravedad ponderadas por los pesos p_i , de modo que, cuando todos los individuos i tienen el mismo peso ($p_i=1$) y la distancia es la euclídea, la inercia de la nube coincide con la varianza total.

ANÁLISIS GENERAL DE LOS MÉTODOS FACTORIALES

Consideramos una tabla rectangular de valores numéricos formada por n filas que representan a n individuos y p columnas que representan a p variables. Los representaremos mediante la matriz X de orden (n,p) y términos x_{ij} (valor que toma la variable j para el individuo i).

$$n \text{ Individuos} \left\{ \begin{array}{c} X = \begin{matrix} & \begin{matrix} p \text{ Variables} \end{matrix} \\ \begin{matrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{matrix} \end{array} \right.$$

Los datos de la tabla anterior pueden representarse en dos espacios distintos. En el **espacio de las variables R^p** se representan los n individuos por sus coordenadas (p -tuplas) o valores que toman para cada una de las p variables. En el **espacio de los individuos R^n** se representan las p variables por sus coordenadas (n -tuplas) o valores que toman para cada uno de los n individuos.

Estos dos espacios están provistos de la distancia euclídea usual. Para dos individuos i e i' , la distancia euclídea entre ellos viene definida como:

$$d(i i') = \sqrt{\sum_j (x_{ij} - x_{i'j})^2}$$

La distancia euclídea entre dos individuos i e i' al cuadrado es la suma de las diferencias existentes entre los valores que toman los individuos para cada variable, elevadas las diferencias al cuadrado para evitar que se compensen las positivas con las negativas.

Si dos individuos que toman valores iguales para todas las variables coinciden en un punto, su distancia es nula. Cuanto mayores sean las diferencias entre los individuos en relación a las variables medidas, más alejadas estarán en el espacio y mayor será su índice de distancias.

Para dos variables j y j' , la distancia euclídea entre ellas viene definida como:

$$d(j \ j') = \sqrt{\sum_i (x_{ij} - x_{ij'})^2}$$

Esta distancia será nula cuando las variables tomen los mismos valores para el conjunto de individuos, y será pequeña cuando estos valores sean próximos; es decir, cuando las variables tengan un comportamiento semejante.

La **cantidad de información de la nube de puntos** se mide por la suma de las distancias desde dichos puntos al origen elevada al cuadrado $\sum_{ij} x_{ij}^2$. Cuando el origen coincide con el centro de gravedad y los pesos sean unitarios, la cantidad de información coincide con la **inercia de la nube** $I_G(N) = \sum_i d^2(i, G)$.

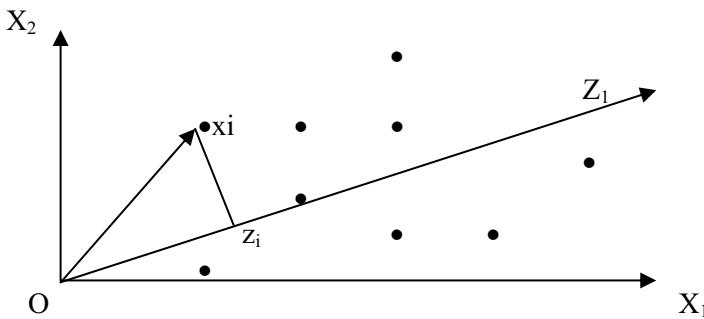
Objetivo general del análisis factorial

Una vez introducido el concepto de información de la nube de puntos, ya podemos especificar el objetivo general del análisis factorial. Este objetivo será buscar un nuevo subespacio de R^p (R^q , $q < p$) que contenga la mayor cantidad posible de información existente en la nube primitiva, y que mejor se ajuste a la nube de puntos y la deforme lo menos posible. El criterio de ajuste es el de los mínimos cuadrados.

Análisis en R^p

Si z_i representa al individuo i en el nuevo subespacio y x_i en el primitivo, se trata de obtener el subespacio que minimice simultáneamente las distancias entre z_i y x_i para todos los puntos de la nube inicial y en proyección, es decir, se trata de obtener el subespacio sobre el cual la nube proyectada se deforme lo menos posible.

En la Figura siguiente se representa la reducción de la nube inicial de puntos x_i a una nube de puntos z_i en un subespacio de dimensión 1 (recta) por el criterio de mínimos cuadrados. Para hacer la reducción a cualquier subespacio de dimensión superior se seguiría un proceso iterativo.



Se trata de minimizar las sumas de los cuadrados de las distancias de x_i a z_i , para evitar que se compensen los valores positivos y negativos. Por lo tanto, se trata de minimizar $\sum_i (\overline{x_i z_i})^2$. Pero, por el teorema de Pitágoras:

$$\overline{Ox_i}^2 = \overline{x_i z_i}^2 + \overline{Oz_i}^2 \Rightarrow \sum_i \overline{Ox_i}^2 = \sum_i \overline{x_i z_i}^2 + \sum_i \overline{Oz_i}^2$$

con lo que se tiene que:

$$\min \sum_i \overline{x_i z_i}^2 = \min \left(\sum_i \overline{Ox_i}^2 - \sum_i \overline{Oz_i}^2 \right) = \max \sum_i \overline{Oz_i}^2$$

Por lo tanto, la minimización de la suma de las distancias al cuadrado $\sum_i (\overline{x_i z_i})^2$ en el espacio original, es equivalente a la maximización de la suma de los cuadrados de las proyecciones $\sum_i \overline{Oz_i}^2$ en el subespacio.

Si u_1 es el vector unitario del eje Z_1 (subespacio de dimensión 1 que mejor ajusta la nube de puntos), la proyección $\overline{Oz_i}$ del punto z_i de la nube inicial sobre el eje Z_1 es el producto escalar de $\overline{Ox_i}$ y u_1 (suma de los productos término a término de los elementos de los vectores $\overline{Ox_i}$ y u_1), es decir, el producto escalar de la fila i -ésima de la matriz X por el vector unitario u_1 , que puede expresarse como $x_i' u_1 = \sum_j x_j u_{1j}$ con $u_1' u_1 = 1$ (por ser u_1 unitario).

Si consideramos las proyecciones $\overline{Oz_i}$ de todos los puntos z_i de la nube inicial sobre el eje Z_1 tenemos que se representan por $X u_1$, siendo su cuadrado $u_1' X' X u_1$.

Por lo tanto para hallar el subespacio de dimensión 1 que mejor ajusta la nube de puntos hay que hallar Z_1 maximizando $u_1'X'Xu_1$, sujeta a la restricción

$$\sum_{j=1}^p u_{1j}^2 = u_1'u_1 = 1.$$

Para resolver este problema de optimización con restricciones se aplica el método de los multiplicadores de Lagrange considerando la función lagrangiana:

$$L(u_1) = u_1'X'Xu_1 - \lambda(u_1'u_1 - 1)$$

Derivando respecto de u_1 e igualando a cero, se tiene:

$$\frac{\partial L}{\partial u_1} = 2X'Xu_1 - 2\lambda u_1 = 0 \Rightarrow (X'X - \lambda I)u_1 = 0$$

Se trata de un sistema homogéneo en u_1 , que sólo tiene solución si el determinante de la matriz de los coeficientes es nulo, es decir, $|X'X - \lambda I| = 0$. Pero la expresión $|X'X - \lambda I| = 0$ es equivalente a decir que λ es un valor propio de la matriz $X'X$.

En general, la ecuación $|X'X - \lambda I| = 0$ tiene p raíces $\lambda_1, \lambda_2, \dots, \lambda_p$, que puedo ordenarlas de mayor a menor $\lambda_1 > \lambda_2 > \dots > \lambda_p$.

En la ecuación $(X'X - \lambda I)u_1 = 0$ podemos multiplicar por u_1' a la derecha, con lo que se tiene $u_1'(X'X - \lambda I)u_1 = 0 \Rightarrow u_1'X'Xu_1 = \lambda$. Por lo tanto, para maximizar $u_1'X'Xu_1$ hay que tomar el mayor valor propio λ de la matriz $X'X$.

Tomando λ_1 como el mayor valor propio de $X'X$ y tomando u_1 como su vector propio asociado normalizado ($u_1'u_1 = 1$), ya tenemos definido el vector director unitario u_1 que define el eje Z_1 (mejor subespacio de dimensión 1 que ajusta la nube de puntos) que vendrá definido como $Z_1 = X u_1$.

Cada individuo tiene una proyección sobre este nuevo eje, y el conjunto de proyecciones se denomina factor, de modo que la nueva variable factor es una combinación lineal de las iniciales, ya que:

$$F_1(i) = x_i' u_1 = \sum_j x_{ij} u_{1j} = x_{i1} u_{11} + \dots + x_{ip} u_{1p}$$

Se trata de una variable artificial que en algún momento se le podrá asignar algún nombre, y otras veces no, pero en todo caso nos permitirá estudiar las relaciones y semejanzas entre los individuos.

La cantidad de información o varianza recogida por el nuevo eje Z_1 es precisamente λ_1 ya que $\lambda_1 = u_1'X'Xu_1 = V(Xu_1) = V(Z_1)$.

La obtención del subespacio de dimensión 2 que mejor ajusta la nube de puntos se hace mediante un proceso iterativo. Una vez hallado el eje (subespacio de dimensión 1) que, pasando por el origen, maximice la suma de cuadrados de las proyecciones sobre él de todos los puntos de la nube inicial, a continuación se busca un segundo eje que, pasando por el origen y siendo perpendicular al primero, maximice la suma de cuadrados de las proyecciones sobre él de todos los puntos de la nube, y así sucesivamente.

Se trata entonces de hallar Z_2 maximizando $V(Z_2) = u_2'X'Xu_2$, sujeta a las restricciones $u_2'u_2=1$ y $u_2'u_1=0$.

Para resolver este problema de optimización con dos restricciones se aplica el método de los multiplicadores de Lagrange considerando la función lagrangiana:

$$L = u_2'X'Xu_2 - \mu(u_2'u_1) - \lambda(u_2'u_2 - 1)$$

Derivando respecto de u_2 e igualando a cero, se tiene:

$$\frac{\partial L}{\partial u_2} = 2X'Xu_2 - \mu u_1 - 2\lambda u_2 = 0$$

Premultiplicando por u_2' tenemos:

$$2u_2'X'Xu_2 - \mu u_2'u_1 - 2\lambda u_2'u_2 = 0$$

Y como $u_2'u_2=1$ y $u_2'u_1=0$, se tiene que $u_2'X'Xu_2=\lambda$, que es el máximo buscado. Si llamamos λ_2 a este máximo ($u_2'X'Xu_2=\lambda_2$) y lo sustituimos en la expresión anterior $2u_2'X'Xu_2 - \mu u_2'u_1 - 2\lambda u_2'u_2 = 0$, se tiene que $X'Xu_2 = \lambda_2 u_2$ (o sea, que u_2 es el vector propio asociado al segundo mayor valor propio λ_2 de $X'X$).

Tomando λ_2 como el segundo mayor valor propio de $X'X$ y tomando u_2 como su vector propio asociado normalizado ($u_2'u_2=1$), ya tenemos definido el vector director del segundo eje Z_2 (que vendrá definido como $Z_2=Xu_2$) perpendicular al primer eje Z_1 , y que permiten hallar el subespacio de dimensión 2 (engendrado por los vectores unitarios u_1 y u_2 directores de Z_1 y Z_2) que mejor ajusta la nube de puntos en el sentido de mínimos cuadrados.

Cada individuo tiene una proyección sobre este nuevo eje Z_2 , proyección que se representa mediante:

$$F_2(i) = \vec{x}_i' \vec{u}_2 = \sum_j x_{ij} u_{2j} = x_{i1} u_{21} + \cdots + x_{ip} u_{2p}$$

Y el conjunto de estas proyecciones (que se denomina factor), constituyen una nueva variable artificial combinación lineal de las p variables iniciales, que es el segundo factor.

De forma similar se obtiene el eje Z_q ($q < p$) perpendicular a todos los anteriores, que se define como $Z_q = X\vec{u}_q$ donde \vec{u}_q es el vector propio de $X'X$ asociado a su q -ésimo mayor valor propio. Suele denominarse también a \vec{u}_q **eje factorial q -ésimo**.

De esta forma se obtiene que el espacio q dimensional ($q < p$) que mejor se ajusta a la nube de puntos está engendrado por los vectores propios $\vec{u}_1, \vec{u}_2, \dots, \vec{u}_q$ asociados a los q mayores valores propios $\lambda_1 > \lambda_2 > \dots > \lambda_q$, de la matriz $X'X$.

Análisis en R^n

Análogamente, en el espacio de los individuos se tratará de buscar los ejes que minimizan la deformación o maximizan la suma de las proyecciones de las variables al cuadrado.

Sea v_1 el vector director del subespacio de dimensión 1 (eje Z'_1) que pasa por el origen. La proyección de un punto j sobre el eje viene dada por $x_j' v_1 = \sum_i x_{ij} v_{1i}$. La proyección de todos los puntos es $X'v_1$ y la suma de sus cuadrados es $v_1' X X' v_1$. Por lo tanto se trata de buscar el vector unitario v_1 que maximice $v_1' X X' v_1$, sujeto a la restricción $v_1' v_1 = 1$. Siguiendo el método utilizado en el caso de R^p , se llega a que v_1 es el vector propio de XX' asociado a su mayor valor propio μ_1 ($XX'v_1 = \mu_1 v_1$).

De esta forma se obtiene que el espacio q dimensional ($q < p$) que mejor se ajusta a la nube de puntos está engendrado por los vectores propios v_1, v_2, \dots, v_q asociados a los q mayores valores propios $\mu_1 > \mu_2 > \dots > \mu_q$, de la matriz $X'X$.

La proyección de un punto j sobre el eje Z'_α ($\alpha = 1, \dots, q$), se representa mediante:

$$G_\alpha(j) = \vec{x}_j' \vec{v}_\alpha = \sum_i x_{ij} u_{\alpha i} = x_{1j} u_{\alpha 1} + \cdots + x_{pj} u_{\alpha n}$$

Y el conjunto de estas proyecciones (que se denomina factor), constituyen una nueva variable artificial combinación lineal de las n variables iniciales, que es el factor α .

Relación entre los análisis en los espacios R^p y R^n

Resulta que valores propios $\mu_1 > \mu_2 > \dots > \mu_q$, asociados a los vectores propios v_1, v_2, \dots, v_q de la matriz XX' son iguales respectivamente a los valores propios $\lambda_1 > \lambda_2 > \dots > \lambda_q$, asociados a los vectores propios u_1, u_2, \dots, u_q de la matriz $X'X$, es decir:

$$\lambda_1 = \mu_1, \lambda_2 = \mu_2, \dots, \lambda_q = \mu_q$$

lo que significa que la cantidad de información o varianza (suma de las proyecciones al cuadrado) recogida por los ejes respectivos en ambos espacios, es la misma.

Para demostrar lo afirmado en el párrafo anterior, partimos de la expresión $XX'v_\alpha = \mu_\alpha v_\alpha$ (que representa el hecho de que v_α es un vector propio de XX' asociado al valor propio μ_α) y premultiplicamos por X' para obtener $(X'X)X'v_\alpha = \mu_\alpha X'v_\alpha$, de donde se deduce que $X'v_\alpha$ es un vector propio de la matriz $X'X$ asociado también al valor propio μ_α . Por lo tanto, a cada vector propio v_α de XX' relativo al valor propio u_α le corresponde un vector propio $X'v_\alpha$ de $X'X$ relativo al mismo valor propio μ_α . Existirá entonces una proporcionalidad entre u_α y $X'v_\alpha$ y todo valor propio no nulo de la matriz XX' es valor propio de la matriz $X'X$. Además, como λ_1 es el mayor valor propio asociado a u_1 , se deduce que $\lambda_1 \geq \mu_1$.

Análogamente en el otro espacio, si partimos de la expresión $X'Xu_\alpha = \lambda_\alpha u_\alpha$ (que representa el hecho de que u_α es un vector propio de $X'X$ asociado al valor propio λ_α) y premultiplicamos por X para obtener $(XX')Xu_\alpha = \lambda_\alpha Xu_\alpha$, de donde se deduce que Xu_α es un vector propio de la matriz XX' asociado también al valor propio λ_α . Por lo tanto, a cada vector propio u_α de $X'X$ relativo al valor propio λ_α le corresponde un vector propio Xu_α de XX' relativo al mismo valor propio λ_α . Existirá entonces una proporcionalidad entre v_α y Xu_α y todo valor propio no nulo de la matriz $X'X$ es valor propio de la matriz XX' . Además, como μ_1 es el mayor valor propio asociado a v_1 , se deduce que $\lambda_1 \leq \mu_1$.

Hemos deducido entonces que $\lambda_1 = \mu_1$, y de igual forma se puede deducir que $\forall \alpha = 1, \dots, q \quad \lambda_\alpha = \mu_\alpha$. Además, conocidos los vectores propios de un subespacio se pueden obtener los del otro sin necesidad de una nueva factorización. Por ejemplo, dados los v_α , la proporcionalidad entre u_α y $X'v_\alpha$ permite escribir $u_\alpha = kX'v_\alpha$. Como $\mu'_\alpha \mu_\alpha = 1$, podemos escribir $k^2 v'_\alpha X'X v_\alpha = 1$. Pero por otra parte, sabemos que $v'_\alpha X'X v_\alpha = \lambda_\alpha$ por ser la suma de las proyecciones al cuadrado (cada valor propio λ_α mide la suma de los cuadrados de las proyecciones sobre el eje α , o sea, $v'_\alpha X'X v_\alpha = \lambda_\alpha$). Esto nos lleva a

$$\text{escribir } k^2 \lambda_\alpha = 1 \Rightarrow k = \frac{1}{\sqrt{\lambda_\alpha}}. \text{ Por lo tanto } u_\alpha = kX'v_\alpha \Rightarrow u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X'v_\alpha.$$

De la misma forma, y partiendo de la proporcionalidad entre v_α y Xu_α se obtiene que

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} u_\alpha$$

Existe entonces una proporcionalidad entre las coordenadas de los puntos individuo sobre el eje factorial α en R^p , Xu_α y las componentes del vector unitario director del eje α en el otro espacio, v_α . Análogamente, las coordenadas de los puntos variables sobre el eje α , $X'v_\alpha$ son proporcionales a las componentes del vector unitario director del eje α en el otro espacio, u_α .

$$G_\alpha = X'v_\alpha = \sqrt{\lambda_\alpha} u_\alpha \quad G_\alpha(j) = \sum_i x_{ij} v_{\alpha i} = \sqrt{\lambda_\alpha} u_{\alpha j}$$

Estas relaciones permiten una reducción de cálculos, de modo que sólo es necesario obtener los valores y vectores propios de una matriz, y a partir de ellos se obtienen los de la otra. Además, por ser iguales los valores propios de las matrices, coinciden las cantidades de información recogida por los dos ejes respectivos en ambos análisis, lo que facilita la superposición de los espacios sobre el mismo gráfico.

Como $\lambda_1 > \lambda_2 > \dots > \lambda_q$, λ_1 debe ser más importante que los demás respecto de la cantidad de información de la nube que recoge. Si a partir del que ocupa el lugar q , los valores propios $\lambda_{q+1}, \lambda_{q+2}, \dots, \lambda_p$ pueden considerarse muy pequeños (próximos a cero), los ejes correspondientes recogen poca información, ya que la suma de las proyecciones al cuadrado sobre esos ejes es pequeña. De esta forma, el conjunto de los q primeros ejes permitirá resumir la nube de puntos con buena precisión. La cantidad de información, tasa de inercia, o parte de la varianza total, recogida por los q primeros ejes factoriales se define mediante $\tau_q = \sum_{h=1}^q \lambda_h / \sum_{h=1}^p \lambda_h$ y mide la parte de dispersión de la nube de puntos recogida en el subespacio R^q .

En general, se define el *porcentaje de inercia explicada por los k primeros ejes factoriales* como:

$$\tau_k = \frac{\sum_{h=1}^k \lambda_h}{\sum_{h=1}^p \lambda_h} = \frac{\sum_{h=1}^k \lambda_h}{\text{traza}(X' X)}$$

Reconstrucción de la tabla inicial de datos a partir de los ejes factoriales

Es posible reconstruir de forma aproximada los valores numéricos de la tabla de datos inicial X a partir de los q primeros ejes, utilizando los vectores directores de los ejes y los valores propios. En efecto:

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} u_\alpha \Rightarrow \sqrt{\lambda_\alpha} v_\alpha = Xu_\alpha \Rightarrow \sqrt{\lambda_\alpha} v_\alpha u_\alpha^\top = Xu_\alpha u_\alpha^\top \Rightarrow \sum_{\alpha=1}^p \sqrt{\lambda_\alpha} v_\alpha u_\alpha^\top = X \sum_{\alpha=1}^p u_\alpha u_\alpha^\top$$

Como los vectores u_α son unitarios y perpendiculares, $\sum_{\alpha=1}^p u_\alpha u_\alpha^\top$ es la matriz identidad, ya que es el producto de la matriz ortogonal de los vectores propios por su traspuesta, que es también su inversa (por ortogonalidad), con lo que:

$$\sum_{\alpha=1}^p \sqrt{\lambda_\alpha} v_\alpha u_\alpha^\top = X \sum_{\alpha=1}^p u_\alpha u_\alpha^\top \Rightarrow X = \sum_{\alpha=1}^p \sqrt{\lambda_\alpha} v_\alpha u_\alpha^\top$$

Si consideramos los q ejes factoriales, se obtiene una representación exacta de la tabla de datos inicial, pero normalmente, a partir del que ocupa el lugar q , los valores propios $\lambda_{q+1}, \lambda_{q+2}, \dots, \lambda_p$ suelen ser muy pequeños (próximos a cero), con lo que los ejes correspondientes recogen poca información, ya que la suma de las proyecciones al cuadrado sobre esos ejes es pequeña. De esta forma estos últimos ejes aportarán poca información a la reconstrucción, considerándose la reconstrucción aproximada de la tabla de datos inicial X dada por:

$$X \approx \sum_{\alpha=1}^q \sqrt{\lambda_\alpha} v_\alpha u_\alpha^\top$$

Se sustituyen así los $n \times p$ números de la matriz X por sólo $n \times q$ números constituidos por q vectores $\sqrt{\lambda_\alpha} v_\alpha$ y los q vectores u_α .

La calidad total de la reconstrucción de la tabla inicial X se mide mediante el coeficiente $\tau_q = \sum_{h=1}^q \lambda_h / \sum_{h=1}^p \lambda_h$

COMPONENTES PRINCIPALES COMO CASO PARTICULAR DEL ANÁLISIS FACTORIAL GENERAL

En el caso en que la tabla de datos de partida esté formada por variables cuantitativas y heterogéneas, es aplicable el análisis factorial general previa tipificación de las variables. El análisis resultante, una vez realizada la tipificación, resulta ser el análisis en componentes principales ya estudiado en un capítulo anterior.

El análisis en componentes principales se utiliza para describir una matriz R de variables continuas del tipo individuos por variables. Es decir, una matriz que recoge el valor que toman cada una de las variables $j, \{j = 1, \dots, p\}$ en cada uno de los individuos u observaciones $i, \{i = 1, \dots, n\}$.

$$n \text{ Individuos} \left\{ \begin{array}{c} \text{\hspace{10em}} p \text{ Variables} \\ R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2j} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ r_{i1} & r_{i2} & \cdots & r_{ij} & \cdots & r_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nj} & \cdots & r_{np} \end{pmatrix} \end{array} \right.$$

Al igual que en el análisis factorial general, los datos de la tabla anterior pueden representarse en dos espacios distintos. En el **espacio de las variables R^p** se representan los n individuos por sus coordenadas (p -tuplas) o valores que toman para cada una de las p variables. En el **espacio de los individuos R^n** se representan las p variables por sus coordenadas (n -tuplas) o valores que toman para cada uno de los n individuos.

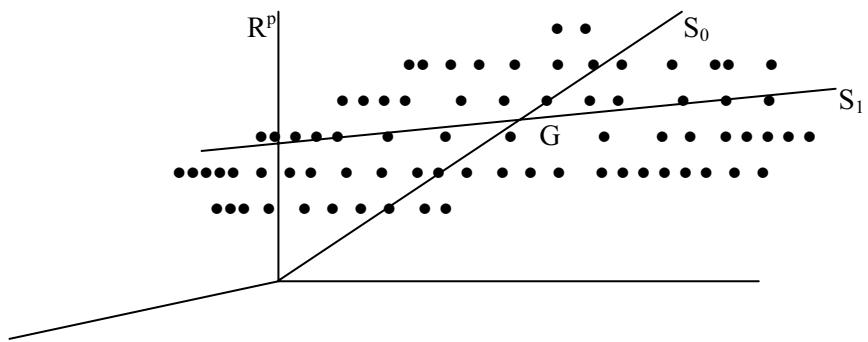
Las variables Figuran en columnas y los individuos, en filas. Éstos pueden ser individuos encuestados, observaciones, marcas, consumidores de un producto, etc. Esta matriz puede ser muy disimétrica, y las variables, muy heterogéneas, tanto en media como en desviación. Por ejemplo, una variable puede medir las ventas en pesetas y otra, tipos de rendimientos, con lo cual las diferencias de medias serían enormes. Por esta razón, antes de aplicar el análisis factorial general a la matriz R, se realiza una transformación de la matriz, como veremos a continuación.

Análisis en R^p

Para evitar que variables que toman valores muy altos tengan un peso muy importante en la determinación de los ejes, se realiza una transformación consistente en centrar los datos de la siguiente forma:

$$x_{ij} = r_{ij} - \bar{r}_j \text{ con } \bar{r}_j = \frac{1}{n} \sum_{i=1}^n r_{ij} = \text{media de la variable } j$$

De esta manera se elimina la influencia del nivel general de las variables, realizándose una traslación del origen al centro de gravedad de la nube. Gráficamente, podemos comprobar la conveniencia de realizar esta operación. Supongamos que la representación de los individuos es la del Gráfico siguiente:



Buscamos el subespacio de dimensión reducida que, pasando por el origen, represente bien la nube de puntos. Si tomamos como solución el subespacio S_0 , no obtendremos una buena representación. Se produce entonces una deformación fuerte al proyectar los puntos individuo sobre S_0 . Sin embargo, lo que tratamos de estudiar no es la posición de los individuos con respecto al origen, sino sus posiciones respectivas, o sea, la forma de la nube. Es evidente que en un caso como el del gráfico que nos ocupa esto se lograría mejor y obtendríamos una representación más fiel sobre el subespacio S_1 , que no pasa por el origen sino por el centro de gravedad G de la nube. Para realizar un análisis general en relación al centro de gravedad G , se traslada el origen de coordenadas al centro de gravedad.

Por otra parte, puede ocurrir que las dispersiones de las distintas variables que forman la tabla de datos sean muy diferentes, lo que hará necesaria otra transformación en los datos de partida, realizando una tipificación como sigue:

$$x_{ij} = \frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}} \quad \text{siendo} \quad s_j^2 = \frac{1}{n} \sum_{i=1}^n (r_{ij} - \bar{r}_j)^2$$

De esta forma, para dos individuos i e i' , la distancia euclídea entre ellos, que en general hemos visto que viene definida por $d(i i') = \sqrt{\sum_j (x_{ij} - x_{i'j})^2}$, puede expresarse como sigue:

$$d^2(i i') = \sum_j (x_{ij} - x_{i'j})^2 = \sum_j \left(\frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}} - \frac{r_{i'j} - \bar{r}_j}{s_j \sqrt{n}} \right)^2 = \frac{1}{n} \sum_j \left(\frac{r_{ij} - r_{i'j}}{s_j} \right)^2$$

De esta forma, todas las variables tendrán una contribución semejante a la determinación de las proximidades y no habrá variables que por ser muy dispersas contribuyan más al cálculo de las distancias.

Otra característica importante de la tipificación realizada lo constituye el hecho de que la matriz de correlaciones C coincida con la matriz $X'X$, cuyo término general $c_{jj'}$ coincide con la correlación entre las variables j y j' , como se muestra en la expresión:

$$c_{jj'} = \sum_i \frac{(r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{s_j \sqrt{n} s_{j'} \sqrt{n}} = \sum_i \frac{(r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{s_j s_{j'} n} = corr(j, j')$$

Por lo tanto, el análisis en componentes principales en R^p consistirá en realizar un análisis factorial general sobre la tabla X tipificada. El análisis consistirá entonces en obtener los vectores propios u_α de la matriz de correlaciones $C=X'X$, y las proyecciones de los individuos sobre los ejes dirigidos por estos vectores propios son las componentes principales, que se obtienen mediante $F_\alpha = X u_\alpha$, donde u_α es el vector propio de $C=X'X$ asociado a su α -ésimo mayor valor propio λ_α . Para el individuo i , su proyección sobre el eje de vector director u_α viene dada por la combinación lineal:

$$F_\alpha(i) = x_i^\top u_\alpha = \sum_j x_{ij} u_{\alpha j} = x_{i1} u_{\alpha 1} + \cdots + x_{ip} u_{\alpha p}$$

La proporción de la variabilidad total recogida por la componente principal h -ésima (**porcentaje de inercia explicada por la componente principal h -ésima**) vendrá dada por:

$$\frac{\lambda_h}{\sum_{h=1}^p \lambda_h} = \frac{\lambda_h}{traza(C)} = \frac{\lambda_h}{p}$$

También se define el **porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales)** como:

$$\frac{\sum_{h=1}^k \lambda_h}{\sum_{h=1}^p \lambda_h} = \frac{\sum_{h=1}^k \lambda_h}{traza(C)} = \frac{\sum_{h=1}^k \lambda_h}{p}$$

Se puede interpretar la nube de individuos en función de los factores, ya que la contribución absoluta de un individuo i a la formación de un eje α ($CTA_\alpha(i)$) es mayor cuanto más alta sea su proyección sobre el eje:

$$CTA_\alpha(i) = \frac{F_\alpha^2(i)}{\sum_i F_\alpha^2(i)}$$

También se puede obtener una medida de la calidad de la representación de un individuo i sobre el eje α a través de la contribución relativa $CTR_\alpha(i)$, o cociente entre la cantidad de información restituida en proyección y la información aportada por i :

$$CTR_\alpha(i) = \frac{F_\alpha^2(i)}{\sum_j x_{ij}^2}$$

Si en su representación en el plano de los factores (individuos representados por sus coordenadas sobre los factores), dos individuos están próximos, pueden interpretarse como individuos de comportamiento semejante, tomando valores próximos para todas las variables medidas sobre ellos.

Análisis en R^n

La transformación realizada en la tabla de datos produce efectos diferentes en este espacio. Así como en R^p se trasladaba el origen al centro de gravedad y se situaba a los individuos alrededor del origen, en R^n la transformación produce una deformación de la nube de puntos. El cambio de escala de cada variable (multiplicación por $1/(s_j \sqrt{n})$) sitúa todos los puntos variables a la distancia 1 del origen. En efecto:

$$d^2(j, O) = \sum_i x_{ij}^2 = \sum_i \left(\frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}} \right)^2 = \sum_i \frac{(r_{ij} - \bar{r}_j)^2 / n}{s_j^2} = \frac{\sum_i (r_{ij} - \bar{r}_j)^2 / n}{s_j^2} = \frac{s_j^2}{s_j^2} = 1$$

Los p puntos están en una hiperesfera de radio 1 cuyo centro es el origen. Al proyectar los puntos sobre el subespacio obtenido al aplicar el análisis factorial general se puede producir una contracción, con lo cual en proyección los puntos estarán situados a una distancia del origen menor o igual a 1.

La distancia entre 2 puntos variables en el espacio R^n puede expresarse en función del coeficiente de correlación $c_{jj'}$ entre las variables j y j' como sigue:

$$\begin{aligned}
 d^2(j, j') &= \sum_i (x_{ij} - x_{ij'})^2 = \sum_i \left(\frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}} - \frac{r_{ij'} - \bar{r}_{j'}}{s_{j'} \sqrt{n}} \right)^2 = \frac{1}{n} \sum_i \left(\frac{r_{ij} - \bar{r}_j}{s_j} - \frac{r_{ij'} - \bar{r}_{j'}}{s_{j'}} \right)^2 \\
 &= \frac{1}{n} \sum_i \left[\frac{(r_{ij} - \bar{r}_j)^2}{s_j^2} - \frac{(r_{ij'} - \bar{r}_{j'})^2}{s_{j'}^2} + 2 \frac{(r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{s_j s_{j'}} \right] = \\
 &= \frac{\frac{1}{n} \sum_i (r_{ij} - \bar{r}_j)^2}{s_j^2} - \frac{\frac{1}{n} \sum_i (r_{ij'} - \bar{r}_{j'})^2}{s_{j'}^2} + 2 \frac{\frac{1}{n} \sum_i (r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{s_j s_{j'}} = \\
 &= \frac{s_j^2}{s_j^2} - \frac{s_{j'}^2}{s_{j'}^2} + 2c_{jj'} = 2(1 - c_{jj'})
 \end{aligned}$$

De esta forma, las proximidades entre los puntos variables se pueden interpretar en términos de correlación, de modo que, si dos variables están muy correlacionadas positivamente ($c_{jj'} \approx 1$), la distancia entre ellas es casi cero ($d^2(j, j') \approx 0$). Si dos variables están muy correlacionadas negativamente ($c_{jj'} \approx -1$), la distancia entre ellas es máxima ($d^2(j, j') \approx 4$). Si las dos variables están incorrelacionadas ($c_{jj'} \approx 0$), la distancia entre ellas es intermedia ($d^2(j, j') \approx 2$).

Para obtener los factores puede no ser necesario diagonalizar la matriz XX' . Como ya se ha visto en el análisis general, los vectores propios de XX' asociados al valor propio λ_α se obtienen a partir de los de la matriz $X'X$ mediante:

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} u_\alpha$$

La proyección de los puntos variables sobre el eje α viene dada por el vector:

$$G_\alpha = X' v_\alpha = \sqrt{\lambda_\alpha} u_\alpha \quad G_\alpha(j) = \sum_i x_{ij} v_{\alpha i} = \sqrt{\lambda_\alpha} u_{\alpha j}$$

Para la variable j , su proyección sobre el eje factorial α también puede expresarse como sigue:

$$G_\alpha(j) = \sum_i x_{ij} v_{\alpha i} = \frac{\sum_i x_{ij} F_\alpha(i)}{\sqrt{\lambda_\alpha}} = \frac{Cov(\alpha, j)}{s_\alpha s_j} = Corr(\alpha, j)$$

Se ha empleado la relación $F_\alpha(i) = v_{\alpha i} \sqrt{\lambda_\alpha}$.

Sobre los planos factoriales los puntos variables están situados en el interior de un círculo de radio unidad centrado en el origen. En efecto, hemos visto que los puntos variables están situados en la hiperesfera de radio unidad (por la transformación de la tipificación realizada en los datos iniciales) y al proyectarlos se puede producir una contracción y acercarse al origen, pero no una dilatación. Cuanto menor sea la pérdida de información, menor será la contracción. Los puntos variables están mejor representados en el plano mientras más próximos estén el borde del círculo. La nube de variables no está centrada en el origen, sino que las variables pueden estar situadas todas al mismo lado del origen si se correlacionan positivamente.

ANÁLISIS FACTORIAL DE CORRESPONDENCIAS

El análisis de correspondencias es un método multivariante factorial de reducción de la dimensión de una tabla de casos-variables con datos cualitativos con el fin de obtener un número reducido de factores, cuya posterior interpretación permitirá un estudio más simple del problema investigado. El hecho de que se manejen variables cualitativas (o, por supuesto, cuantitativas categorizadas) confiere a esta prueba factorial una característica diferencial: No se utilizan como datos de partida mediciones individuales, sino frecuencias de una tabla; es decir, número de individuos contenidos en cada casilla. El análisis factorial es de aplicación incluso con sólo dos caracteres o variables cualitativas (**análisis de correspondencias simple**), cada una de las cuales puede presentar varias modalidades o categorías. El método se generaliza cuando el número de variables o caracteres cualitativos es mayor de dos (**análisis de correspondencias múltiple**).

El conocido tratamiento conjunto de dos caracteres o variables cualitativas a través de la prueba de asociación o independencia de la χ^2 proporcionaba exclusivamente información sobre la relación significativa o no entre ambas, sin aclarar qué categorías o modalidades estaban implicadas. Sin embargo, el análisis de correspondencias extrae relaciones entre categorías y define similaridades o disimilaridades entre ellas, lo que permitirá su agrupamiento si se detecta que se corresponden. Y todo esto queda plasmado en un espacio dimensional de escasas variables sintéticas o factores que pueden ser interpretados o nombrados y que, además, deben condensar el máximo posible de información. Representaciones gráficas o mapas de correspondencias permiten visualizar globalmente las relaciones obtenidas.

Por obedecer a la sistemática general del análisis factorial, las dimensiones que definen el espacio en que se representan las categorías se obtienen como factores cuantitativos, por lo que el análisis de correspondencias acaba siendo un método de extracción de variables ficticias cuantitativas a partir de variables cualitativas originales, al definir aquéllas las relaciones entre las categorías de éstas. Esto puede permitir la aplicación posterior de otras pruebas multivariantes cuantitativas (regresión, clusters...). Una posibilidad propia de este análisis es la inclusión a posteriori de una nueva categoría de alguna de las variables (categoría suplementaria) que, no habiendo participado en el cálculo, interese representar para su comparación con las originales. La abundancia y vistosidad de los resultados obtenidos hacen de esta prueba una magnífica fuente de hipótesis de trabajo para continuar la investigación.

El carácter cualitativo de las variables también obliga a un proceso metodológico distinto. Si se trata de estudios de similaridad o disimilaridad entre categorías, se habrá de cuantificar la diferencia o distancia entre ellas. En una tabla de frecuencias cada categoría de una variable está formada por un conjunto de individuos distribuidos en cada una de las categorías de la otra. Por tanto, el proceso para hallar la distancia entre dos categorías de una variable es el utilizado en Estadística para el cálculo del desajuste de dos distribuciones, por medio de las diferencias (desajustes) cuadráticas (para evitar enjuagar diferencias positivas con negativas) relativas (es menos clara una diferencia de dos individuos en cuatro que en un dos por ciento). La suma de estas diferencias cuadráticas relativas entre las frecuencias de ambas distribuciones no es otra cosa que el conocido concepto de la χ^2 . Así, el análisis de correspondencias puede considerarse como un análisis de componentes principales aplicado a variables cualitativas que, al no poder utilizar correlaciones, se basa en la distancia no euclídea de la χ^2 .

ANÁLISIS DE CORRESPONDENCIAS SIMPLE

Ya sabemos que el análisis factorial de correspondencias simple está particularmente adaptado para tratar tablas de contingencia, representando los efectivos existentes en las múltiples modalidades (categorías) combinadas de dos caracteres (variables cualitativas). Si cruzamos en una tabla de contingencia el carácter I con modalidades desde $i=1$ hasta $i=n$ (en filas), con el carácter J con modalidades desde $j=1$ hasta $j=p$ (en columnas), podemos representar el número de unidades estadísticas que pertenecen simultáneamente a la modalidad i del carácter I y a la modalidad j del carácter J mediante k_{ij} . En este caso, la distinción entre observaciones y variables en el cuadro de doble entrada es artificial, pero, por similitud con componentes principales, suele hablarse a veces de individuos u observaciones cuando nos referimos al conjunto de las modalidades del carácter I (filas), y de variables cuando nos referimos al conjunto de las modalidades del carácter J (columnas), tal y como se observa en la Tabla siguiente:

<i>I</i>	<i>J</i>	1	2	...	<i>j</i>	...	<i>p</i>
1							
2					⋮		
⋮					...	<i>k_{ij}</i>	...
<i>i</i>						⋮	
⋮						⋮	
<i>n</i>							

De una forma general puede considerarse que los objetivos que se persiguen cuando se aplica el análisis factorial de correspondencias son similares a los perseguidos con la aplicación del análisis de componentes principales, y pueden resumirse en los dos puntos siguientes:

- Estudio de las relaciones existentes en el interior del conjunto de modalidades del carácter I y estudio de las relaciones existentes en el interior del conjunto de modalidades del carácter J.
- Estudio de las relaciones existentes entre las modalidades del carácter I y las modalidades del carácter J.

La tabla de datos (k_{ij}) es una matriz K de orden (n, p) donde k_{ij} representa la frecuencia absoluta de asociaciones entre los elementos i y j , es decir el número de veces que se presentan simultáneamente las modalidades i y j de los caracteres I y J.

Utilizaremos la siguiente notación:

$$k_{i \cdot} = \sum_{j=1}^p k_{ij} = \text{efectivo total de la fila } i.$$

$$k_{\cdot j} = \sum_{i=1}^n k_{ij} = \text{efectivo total de la columna } j.$$

$$k_{\cdot \cdot} = \sum_{i=1}^n \sum_{j=1}^p k_{ij} = \text{efectivo total de la población.}$$

El método buscado para el análisis factorial de correspondencias simple deberá ser simétrico con relación a las líneas y columnas de K (para estudiar las relaciones en el interior de los conjuntos I y J) y deberá permitir comparar las distribuciones de frecuencias de las dos características (para estudiar las relaciones entre los conjuntos I y J).

Para comparar dos líneas entre sí (filas o columnas) en una tabla de contingencia, no interesan los valores brutos sino los porcentajes o distribuciones condicionadas. En una tabla de contingencia, el análisis buscado debe trabajar no con los valores brutos k_{ij} sino con **perfiles** o porcentajes. No interesa poner de manifiesto las diferencias absolutas que existen entre dos líneas, sino que los elementos i, i' (j, j') se consideran semejantes si presentan la misma distribución condicionada.

Una primera caracterización de las modalidades i del carácter I (variables i) puede hacerse a partir del peso relativo (expresado en tanto por uno) de cada modalidad del carácter J en la modalidad i , $\frac{k_{i1}}{k_{i.}}, \frac{k_{i2}}{k_{i.}}, \dots, \frac{k_{ip}}{k_{i.}}$, que denominamos

perfil de la variable i , y que es la distribución de frecuencias condicionada del carácter J para $I=i$.

De modo análogo la caracterización de las modalidades j del carácter J (observaciones j) puede hacerse a partir del peso relativo (expresado en tanto por uno) de cada modalidad del carácter I en la modalidad j , $\frac{k_{1j}}{k_{.j}}, \frac{k_{2j}}{k_{.j}}, \dots, \frac{k_{nj}}{k_{.j}}$, que denominamos **perfil de la observación j** , y que es la distribución de frecuencias condicionada del carácter I para $J=j$.

Formación de las nubes y definición de distancias

En R^p tomaremos la nube de n puntos i (n filas de la tabla de perfiles de las variables i) cuyas coordenadas son $\frac{k_{i1}}{k_{i.}}, \frac{k_{i2}}{k_{i.}}, \dots, \frac{k_{ip}}{k_{i.}}$ $i=1\dots n$

En R^n se forma la nube de p puntos j (p columnas de la tabla de perfiles de las observaciones j) cuyas coordenadas son $\frac{k_{1j}}{k_{.j}}, \frac{k_{2j}}{k_{.j}}, \dots, \frac{k_{nj}}{k_{.j}}$ $j=1\dots p$

Las transformaciones realizadas son idénticas en los dos espacios R^p y R^n . Sin embargo, ello va a llevar a transformaciones analíticas diferentes. Los nuevos datos en R^n no son la traspuesta de la matriz en R^p . Esto nos conduce a *realizar dos análisis factoriales diferentes, uno en cada espacio*. Pero encontraremos unas relaciones entre los factores que permitirán reducir los cálculos a una sola factorización facilitando además la interpretación.

A partir de ahora se trabajará con la *tabla de contingencia en frecuencias relativas* $f_{ij} = \frac{k_{ij}}{k}$ con $k = \sum_{i=1}^n \sum_{j=1}^p k_{ij}$. Tendremos el siguiente esquema:

<i>Perfil de las líneas en R^p</i>					
	1	...	j	...	p
1					
:					
i			k_{ij}		
:					
n					

→

<i>Perfil de las líneas en R^p</i>					
	1	...	j	...	p
1					
:					
i			$f_{ij} / f_{i\cdot}$		
:					
n					



<i>Perfil de las columnas en R^n</i>					
	1	...	j	...	p
1					
:					
i			$f_{ij} / f_{\cdot j}$		
:					
n					

$f_{i\cdot} = \frac{k_{i\cdot}}{k}$ $f_{\cdot j} = \frac{k_{\cdot j}}{k}$

$\frac{k_{ij}}{k_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}}$ $\frac{k_{ij}}{k_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}$

El análisis factorial de correspondencias trabaja con perfiles, pero no olvida las diferencias entre los efectivos de cada línea o columna, sino que les asigna un peso proporcional a su importancia en el total. En R^p cada punto i está afectado por un peso $f_{i\cdot}$ y en R^n cada punto j está afectado por un peso $f_{\cdot j}$ con lo que, de esta forma, se evita que al trabajar con perfiles se privilegie a las clases de efectivos pequeños.

El hecho de trabajar con perfiles, en vez de con los valores absolutos iniciales nos lleva a utilizar la distancia ji-cuadrado (distancia entre distribuciones) en vez de la euclídea. Partiendo de la definición de distancia ji-cuadrado dada al principio del Capítulo, en el análisis de correspondencias la distancia entre los individuos (puntos fila) i e i' en R^p vendrá definida como:

$$d_{ii'}^2 = \sum_{j=1}^p \frac{1}{k_{\cdot j}/k} \left(\frac{k_{ij}}{k_{\cdot j}} - \frac{k_{i'j}}{k_{\cdot j}} \right)^2 = \sum_{j=1}^p \frac{1}{k_{\cdot j}/k} \left(\frac{k_{ij}/k}{k_{\cdot j}/k} - \frac{k_{i'j}/k}{k_{\cdot j}/k} \right)^2 = \sum_{j=1}^p \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{i'j}}{f_{\cdot j}} \right)^2$$

De forma similar, en el análisis de correspondencias la distancia entre las variables (puntos columna) j y j' en R^n vendrá definida como:

$$d_{jj'}^2 = \sum_{i=1}^n \frac{1}{k_{i\cdot}/k} \left(\frac{k_{ij}}{k_{i\cdot}} - \frac{k_{ij'}}{k_{i\cdot}} \right)^2 = \sum_{i=1}^n \frac{1}{k_{i\cdot}/k} \left(\frac{k_{ij}/k}{k_{i\cdot}/k} - \frac{k_{ij'}/k}{k_{i\cdot}/k} \right)^2 = \sum_{i=1}^n \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{ij'}}{f_{i\cdot}} \right)^2$$

Realmente la única diferencia entre esta distancia y la euclídea es la ponderación, lo que evita que pequeñas diferencias entre las componentes de la líneas influyan mucho en la distancia. El uso de la distancia ji-cuadrado estabiliza los datos, hasta el punto de que, por el principio de la equivalencia distribucional, dos líneas (filas o columnas) con el mismo perfil pueden ser sustituidas por una sola afectada por una masa igual a la suma de las masas, sin que se alteren las distancias entre los demás pares de puntos en R^p o R^n .

Ejes factoriales: Análisis en R^p

Como el análisis es simétrico para filas y columnas, en el análisis factorial de correspondencias suele elegirse para columnas la dimensión más pequeña ($p < n$).

En R^p el objetivo es obtener una representación simplificada de los puntos fila cuyas coordenadas son $f_{ij}/f_{i\cdot}$, $j=1,\dots,p$. Estos puntos están afectados de un peso o masa $f_{i\cdot}$ y la distancia entre ellos se mide a través de la distancia ji-cuadrado. Vamos a ver que este análisis de correspondencias es equivalente a un análisis en componentes principales de una tabla deducida de la inicial. Tenemos que la distancia entre los individuos (puntos fila) i e i' en R^p puede transformarse como sigue:

$$d_{ii'}^2 = \sum_{j=1}^p \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 = \sum_{j=1}^p \left(\frac{f_{ij}}{f_{i\cdot} \sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot} \sqrt{f_{\cdot j}}} \right)^2$$

expresión que representa la distancia euclídea entre los puntos de coordenadas

$$\frac{f_{ij}}{f_{i\cdot} \sqrt{f_{\cdot j}}} \text{ y } \frac{f_{i'j}}{f_{i'\cdot} \sqrt{f_{\cdot j}}}.$$

Por lo tanto, realizar un análisis con la distancia ji-cuadrado en la tabla f_{ij}/f_i , es equivalente al realizar un análisis con la distancia euclídea en la tabla $\frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}}$ con los pesos f_i .

Las coordenadas del centro de gravedad de la nube de puntos $\frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}}$ con los pesos f_i son:

$$g_j = \sum_{i=1}^n \frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}} f_i = \sum_{i=1}^n \frac{f_{ij}}{\sqrt{f_{\cdot j}}} = \frac{f_{\cdot j}}{\sqrt{f_{\cdot j}}} = \sqrt{f_{\cdot j}}$$

Como el análisis en componentes principales es centrado, trasladaremos el origen al centro de gravedad, con lo que las coordenadas de la nube de puntos $\frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}}$ pasarán a ser $\frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}}$

La inercia de la nube de puntos pasará a ser:

$$I = \sum_{i=1}^n f_i \cdot d^2(i, G) = \sum_{i=1}^n f_i \cdot \sum_{j=1}^p \left(\frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right)^2 = \sum_{i,j}^{n,p} \frac{(f_{ij} - f_i \cdot f_{\cdot j})^2}{f_i \cdot f_{\cdot j}}$$

La proyección de un punto sobre un nuevo eje de vector unitario u_1 viene dada por el producto escalar del punto y el vector u_1 , es decir:

$$F_1(i) = \sum_{j=1}^p \left(\frac{f_{ij}}{f_i \cdot \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) u_{1j}$$

Para hallar el primer factor, se trata de buscar u_1 que maximice la inercia de la nube proyectada, es decir, la suma de los cuadrados de las proyecciones cada una multiplicada por su peso ($\max \sum_{i=1}^n f_i \cdot F_1^2(i)$). Pero sabemos que este problema es equivalente a diagonalizar (vectores propios) la matriz Z de término general:

$$z_{jj'} = \sum_{i=1}^n f_i \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{f_{ij'}}{f_{i \cdot} \sqrt{f_{\cdot j'}}} - \sqrt{f_{\cdot j'}} \right) = \sum_{i=1}^n \left(\frac{f_{ij} - f_{i \cdot} f_{\cdot j}}{\sqrt{f_{i \cdot} \sqrt{f_{\cdot j}}}} \right) \left(\frac{f_{ij'} - f_{i \cdot} f_{\cdot j'}}{\sqrt{f_{i \cdot} \sqrt{f_{\cdot j'}}}} \right)$$

Esta matriz se puede expresar como $Z=X'X$ siendo X la matriz de término general $x_{ij} = \frac{f_{ij} - f_{i \cdot} f_{\cdot j}}{\sqrt{f_{i \cdot} \sqrt{f_{\cdot j}}}}$. Por lo tanto, el análisis factorial de correspondencias relativo a la tabla inicial k_{ij} , es equivalente al análisis en componentes principales para la matriz de término general x_{ij} .

De todas formas, se pueden realizar algunas simplificaciones, basadas en el hecho de que el vector u_p director del eje p de coordenadas $(\sqrt{f_{1 \cdot}}, \sqrt{f_{2 \cdot}}, \dots, \sqrt{f_{p \cdot}})$ es un vector propio de $Z=X'X$ asociado al valor propio 0, ya que partiendo de la expresión desarrollada de $Z u_p$ tenemos:

$$\begin{aligned} \sum_{j'=1}^p \sum_{i=1}^n f_i \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{f_{ij'}}{f_{i \cdot} \sqrt{f_{\cdot j'}}} - \sqrt{f_{\cdot j'}} \right) \sqrt{f_{\cdot j'}} &= \sum_{i=1}^n f_i \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{\sum_{j'=1}^p f_{ij'}}{\sum_{j'=1}^p f_{\cdot j'}} - \frac{\sum_{j'=1}^p f_{\cdot j'}}{\sum_{j'=1}^p f_{\cdot j'}} \right) = \\ \sum_{i=1}^n f_i \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) \left(\frac{f_{i \cdot}}{f_{i \cdot}} - 1 \right) &= 0 \Rightarrow \forall j \quad \sum_{j'=1}^p z_{jj'} \sqrt{f_{\cdot j'}} = 0 \Rightarrow Z u_p = 0 u_p \end{aligned}$$

Los restantes vectores propios de Z deben ser ortogonales a u_p , luego:

$$\sum_{j=1}^p u_{\alpha j} \sqrt{f_{\cdot j}} = 0$$

con lo que todos los vectores propios de $Z=X'X$, $\forall \alpha \neq p$ son también vectores propios de $S=X^{*'}X^*$ siendo $x_{ij}^{*} = \frac{f_{ij}}{\sqrt{f_{i \cdot} \sqrt{f_{\cdot j}}}}$ ya que $\sum_{j'=1}^p z_{jj'} u_{\alpha j'} = \sum_{j'=1}^p s_{jj'} u_{\alpha j'} \quad \forall \alpha \neq p$

El vector u_p es también vector propio de S , pero asociado al valor propio 1, por lo que el análisis puede realizarse sobre la tabla X^* no centrada. Esto conlleva que la proyección del punto i sobre el eje α toma la expresión:

$$F_\alpha(i) = \sum_{j=1}^p \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} - \sqrt{f_{\cdot j}} \right) u_{\alpha j} = \sum_{j=1}^p \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} \right) u_{\alpha j}$$

Ejes factoriales: Análisis en R^n

Como el análisis es simétrico para filas y columnas, se pueden deducir rápidamente los resultados para el análisis en R^n . Así tendremos que las coordenadas de los puntos j serán f_{ij}/f_j , su peso será f_j , el centro de gravedad G tendrá de coordenadas $g_i = \sqrt{f_i}$, la proyección de un punto j sobre el eje α cuyo vector director

$$\text{es } v_\alpha \text{ es } G_\alpha(j) = \sum_{i=1}^n \left(\frac{f_{ij}}{\sqrt{f_i}} - \sqrt{f_i} \right) v_{\alpha i}, \text{ la matriz a diagonalizar es } W \text{ donde}$$

$$w_{ii'} = \sum_{j=1}^p f_{\cdot j} \left(\frac{f_{ij}}{\sqrt{f_i}} - \sqrt{f_i} \right) \left(\frac{f_{i'j}}{\sqrt{f_{i'}}} - \sqrt{f_{i'}} \right) = \sum_{j=1}^p \left(\frac{f_{ij} - f_{\cdot j} f_{\cdot j}}{\sqrt{f_i} \sqrt{f_{\cdot j}}} \right) \left(\frac{f_{i'j} - f_{\cdot j} f_{\cdot j}}{\sqrt{f_{i'}} \sqrt{f_{\cdot j}}} \right)$$

Además el vector v_p director del eje p de coordenadas $(\sqrt{f_1}, \sqrt{f_2}, \dots, \sqrt{f_n})$ es un vector propio de $W=XX'$ asociado al valor propio 0, y todos los vectores propios v_α de $W=XX'$ $\forall \alpha \neq p$ son también vectores propios de $W^*=X^*X^{**}$ siendo v_p el vector propio asociado al valor propio 1. Esto conlleva a que la proyección del punto j sobre el eje α toma la expresión:

$$G_\alpha(j) = \sum_{i=1}^n \left(\frac{f_{ij}}{\sqrt{f_i}} - \sqrt{f_i} \right) v_{\alpha i} = \sum_{i=1}^n \left(\frac{f_{ij}}{\sqrt{f_i}} \right) v_{\alpha i}$$

Relación entre los análisis en R^p y R^n

Los valores propios λ_α no nulos de las matrices $X'X$ y XX' son los mismos. Además los vectores propios u_α de $X'X$ y vectores propios v_α de XX' están relacionados mediante las expresiones $v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X u_\alpha$ y $u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X' v_\alpha$, y sustituyendo los términos de X por sus valores en función de las frecuencias en el análisis de correspondencias se tiene lo siguiente:

$$v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^p \frac{f_{ij}}{\sqrt{f_i} \sqrt{f_{\cdot j}}} u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) \sqrt{f_i} \Rightarrow F_\alpha(i) = \frac{\sqrt{\lambda_\alpha} v_{\alpha i}}{\sqrt{f_i}}$$

$$u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^n \frac{f_{ij}}{\sqrt{f_i} \sqrt{f_{\cdot j}}} v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} G_\alpha(j) \sqrt{f_{\cdot j}} \Rightarrow G_\alpha(j) = \frac{\sqrt{\lambda_\alpha} u_{\alpha j}}{\sqrt{f_{\cdot j}}}$$

Estas relaciones entre los dos subespacios permiten representar simultáneamente los puntos línea y los puntos columna sobre los mismos gráficos, lo que favorece la interpretación de los resultados. Tenemos lo siguiente:

- La proyección de los puntos j sobre el eje α puede expresarse en función de la proyección de los puntos i (utilizando que $v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) \sqrt{f_{i \cdot}}$) como sigue:

$$G_\alpha(j) = \sum_{i=1}^n \left(\frac{f_{ij}}{f_{\cdot j} \sqrt{f_{i \cdot}}} \right) v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^n \left(\frac{f_{ij}}{f_{\cdot j} \sqrt{f_{i \cdot}}} \right) F_\alpha(i) \sqrt{f_{i \cdot}} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^n \left(\frac{f_{ij}}{f_{\cdot j}} \right) F_\alpha(i)$$

- La proyección de los puntos i sobre el eje α puede expresarse en función de la proyección de los puntos j (utilizando que $u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} G_\alpha(j) \sqrt{f_{\cdot j}}$) como sigue:

$$F_\alpha(i) = \sum_{j=1}^p \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} \right) u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^p \left(\frac{f_{ij}}{f_{i \cdot} \sqrt{f_{\cdot j}}} \right) G_\alpha(j) \sqrt{f_{\cdot j}} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^p \left(\frac{f_{ij}}{f_{i \cdot}} \right) G_\alpha(j)$$

Según las expresiones anteriores resultan las relaciones siguientes:

- La proyección de un punto i sobre el eje α , $F_\alpha(i)$, es el baricentro (salvo el coeficiente $1/\sqrt{\lambda_\alpha}$) de las proyecciones de los puntos j sobre el mismo eje, cada punto afectado del peso $f_{ij}/f_{i \cdot}$ que es su importancia relativa en i .
- La proyección de un punto j sobre el eje α , $G_\alpha(j)$, es el baricentro (salvo el coeficiente $1/\sqrt{\lambda_\alpha}$) de las proyecciones de los puntos i sobre el mismo eje, cada punto afectado del peso $f_{ij}/f_{\cdot j}$ que es su importancia relativa en j .

Las relaciones anteriores, llamadas *relaciones baricéntricas, permiten pasar de un espacio a otro y representar simultáneamente sobre el mismo plano los puntos fila y columna, permitiendo así clarificar las relaciones entre filas y columnas.*

Reconstrucción de la tabla de frecuencias

En el análisis general habíamos visto que se reconstruía la tabla de frecuencias inicial a partir de los factores mediante $X = \sum_{\alpha=1}^p \sqrt{\lambda_\alpha} v_\alpha u_\alpha^\top$.

Si en la expresión anterior sustituimos $u_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} G_\alpha(j) \sqrt{f_{\cdot j}}$ y también

$v_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) \sqrt{f_{i \cdot}}$, se tiene lo siguiente:

$$\frac{f_{ij}}{\sqrt{f_{\cdot j}} \sqrt{f_{i \cdot}}} = \sum_{\alpha=1}^p \sqrt{\lambda_\alpha} \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) \sqrt{f_{i \cdot}} \frac{1}{\sqrt{\lambda_\alpha}} G_\alpha(j) \sqrt{f_{\cdot j}} \Rightarrow \frac{f_{ij}}{f_{i \cdot} f_{\cdot j}} = \sum_{\alpha=1}^p \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) G_\alpha(j)$$

pero $\lambda_1=1$, $u_{1j}=\sqrt{f_{\cdot j}}$, $v_{1i}=\sqrt{f_{i \cdot}}$ y $F_\alpha(i)=v_i \sqrt{\lambda_\alpha}/\sqrt{f_{i \cdot}}$. $\Rightarrow F_1(i)=\sqrt{f_{i \cdot}}/\sqrt{f_{i \cdot}}=1$ y $G_1(j)=\sqrt{f_{\cdot j}}/\sqrt{f_{\cdot j}}=1$, con lo que podemos reconstruir la tabla de frecuencias mediante:

$$f_{ij} = f_{i \cdot} f_{\cdot j} \left[1 + \sum_{\alpha=2}^p \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) G_\alpha(j) \right]$$

ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES

Hemos visto que el análisis factorial de correspondencias es de aplicación con dos caracteres o variables cualitativas (*análisis de correspondencias simple* o sencillamente *análisis factorial de correspondencias*), cada una de las cuales puede presentar varias modalidades o categorías. Pero el método es generalizable al caso de un número de variables o caracteres cualitativos mayor de dos (*análisis de correspondencias múltiple*).

Ya sabemos que el análisis factorial de correspondencias simple está particularmente adaptado para tratar tablas de contingencia, representando los efectivos existentes en las múltiples modalidades (categorías) combinadas de dos caracteres (variables cualitativas). Si cruzamos en una tabla de contingencia el carácter I con modalidades desde $i=1$ hasta $i=n$ (en filas), con el carácter J con modalidades desde $j=1$ hasta $j=p$ (en columnas), podemos representar el número de unidades estadísticas que pertenecen simultáneamente a la modalidad i del carácter I y a la modalidad j del carácter J mediante k_{ij} . En este caso, suele hablarse a veces de individuos u observaciones cuando nos referimos al conjunto de las modalidades del carácter I (filas), y de variables cuando nos referimos al conjunto de las modalidades del carácter J (columnas), pero sólo por simple similitud con el análisis en componentes principales, ya que los resultados hemos visto que son totalmente simétricos. Cuando el número de caracteres es mayor que dos (en vez de tener sólo los caracteres I, J tenemos los caracteres J_1, J_2, \dots, J_Q) ya no se puede hablar de tabla de contingencia y la representación tabulada de los datos se complica. No obstante, el análisis en correspondencias múltiples permite estudiar las relaciones entre las modalidades de todas las características cualitativas consideradas.

En el análisis de correspondencias múltiples se ordenan los datos en una tabla Z denominada **tabla disyuntiva completa** que consta de un conjunto de individuos $I=1,\dots,i,\dots,n$ (en filas), un conjunto de variables o caracteres cualitativos J_1,\dots,J_k,\dots,J_Q (en columnas) y un conjunto de modalidades excluyentes $1,\dots,m_k$ para cada carácter cualitativo. El número total de modalidades será entonces $J=\sum_{k=1}^Q m_k$. La tabla disyuntiva completa Z de dimensión $I \times J$ tiene el siguiente aspecto:

		J		
		$\leftarrow J_1 \rightarrow$	$\leftarrow J_k \rightarrow$	$\leftarrow J_Q \rightarrow$
		1..... m_1	1..... m_k	1..... m_Q
I	1	Z_I	Z_k
.
i				Z_Q
.	.			
n				

$Z = ZI \dots Zk \dots ZQ$

El elemento z_{ij} de la tabla toma el valor 0 o 1 según que el individuo i haya elegido (esté afectado por) la modalidad j o no. Por lo tanto, cada rectángulo de la tabla disyuntiva completa puede considerarse, aunque no lo sea, como una tabla de contingencia cuyos elementos son 0 o 1. La tabla disyuntiva completa Z consta entonces de Q subtablas yuxtapuestas, con la finalidad de obtener una representación simultánea de todas las modalidades (columnas) de todos los individuos (filas). Si las modalidades son excluyentes, cada subtabla tiene un único 1 en cada una de sus filas.

Si conservamos la notación que hemos manejado hasta ahora tenemos:

$$z_{ij} = k_{ij} = 0 \text{ ó } 1.$$

$$k_{i\cdot} = \sum_j k_{ij} = Q = \text{número de modalidades} \quad (\text{cada subtabla tiene un único 1 en cada fila}).$$

$$k_{\cdot j} = \sum_i k_{ij} = \text{número de individuos que poseen la modalidad } j.$$

$$f_{ij}/f_{i\cdot} = k_{ij}/k_{i\cdot} = 1/Q = \text{inverso del número de modalidades} \quad (0 \text{ si el individuo no elige } j).$$

Obtención de los factores: Tabla de Burt

Para obtener los factores es necesario diagonalizar la matriz $V=D^{-1}B/Q$ donde $B=Z'Z$ es la tabla de Burtz, matriz simétrica formada por Q^2 bloques, de modo que sus bloques de la diagonal $Z'kZ_k$ son tablas diagonales que cruzan una variable con ella misma, siendo los elementos de la diagonal los efectivos de cada modalidad $k_{\cdot j}$. Los bloques fuera de la diagonal son tablas de contingencia obtenidas cruzando las características de dos en dos $Z'kZ_k$ cuyos elementos son las frecuencias de asociación de las dos modalidades correspondientes. La matriz D es una matriz diagonal cuyos elementos diagonales son los de la matriz de Burtz, siendo nulos el resto de los elementos. El aspecto de la tabla de Burt es el siguiente:

	J_1	J_2	\dots	J_Q
J_1	$0 \ddots 0$	C_{12}	\dots	C_{1Q}
J_2	C_{21}	$0 \ddots 0$	\dots	C_{2Q}
\vdots	\vdots	\vdots	\ddots	\vdots
J_Q	C_{Q1}	C_{Q2}	\dots	$0 \ddots 0$

Las fórmulas de transición que **permiten representar simultáneamente los puntos línea y los puntos columna sobre los mismos gráficos relacionando así los resultados en los dos subespacios** tomarán ahora las siguientes expresiones:

$$F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^p \left(\frac{f_{ij}}{f_{\cdot j}} \right) G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{Q} \sum_{j=1}^p k_{ij} G_\alpha(j)$$

$$G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^n \left(\frac{f_{ij}}{f_{\cdot j}} \right) F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{k_{\cdot j}} \sum_{i=1}^n k_{ij} F_\alpha(i)$$

Si tenemos en cuenta que $k_{ij}=1$ cuando el individuo i posee la modalidad j y cero cuando no, la **proyección de un punto individuo i sobre el eje α** , $F_\alpha(i)$, es el baricentro (salvo un coeficiente de dilatación $1/\sqrt{\lambda_\alpha}$) de las proyecciones de los puntos modalidades sobre el eje, $G_\alpha(j)$. Todas las modalidades están afectadas del mismo peso $1/Q$. Análogamente, la **proyección de un punto modalidad j sobre el eje α** , $G_\alpha(j)$, es el baricentro (salvo un coeficiente de dilatación $1/\sqrt{\lambda_\alpha}$) de las proyecciones de los puntos individuos que poseen esa modalidad sobre el eje, $F_\alpha(i)$, todos ellos afectados del mismo peso $k_{\cdot j}$.

El **centro de gravedad de la nube de puntos variables** $N(j)$ en Análisis Factorial de Correspondencias (ACM) es $\sqrt{f_{\cdot j}}$, que en este caso puede equipararse a una distribución uniforme $1/\sqrt{n}$, ya que $k_{\cdot j} = \sum_j k_{ij} = Q \Rightarrow \sum_i k_{\cdot j} = nQ \Rightarrow f_{\cdot j} = 1/n$.

El **centro de gravedad de las modalidades de cada variable**, cada una ponderada por su peso, es el mismo que el de la nube de modalidades N(J), es decir, $1/\sqrt{n}$, ya que el centro de gravedad de la subtabla $I \times J_k$ se obtiene a partir de su distribución marginal. Como sólo recoge una variable, la suma de cada línea es 1 y el total de la tabla es n , de donde $f_i = 1/n$.

Como el Análisis Factorial de Correspondencias es centrado y el centro de gravedad de las modalidades de una variable coincide con el del conjunto J, y con el origen, las modalidades de cada variable están centradas en torno al origen, no pudiendo tener todas el mismo signo.

Al igual que en cualquier Análisis Factorial de Correspondencias, se calculan las **ayudas a la interpretación para cada fila y columna**, definiendo la contribución de una variable J_k al factor α , como la suma de las contribuciones de las modalidades de la variable:

$$CTA_\alpha(J_k) = \sum_{j \in J_k} CTA_\alpha(j)$$

La parte de inercia debida a una modalidad j es mayor cuanto menor sea el efectivo de esa modalidad. Si G representa el centro de gravedad, la **inercia debida a la modalidad j** viene dada por:

$$I(j) = f_{\cdot j} \cdot d^2(G, j) = f_{\cdot j} \sum_{i=1}^n \left(\frac{f_{ij}}{f_{\cdot j} \sqrt{f_{\cdot i}}} - \sqrt{f_{\cdot i}} \right)^2 = \frac{k_{\cdot j}}{nQ} \sum_{i=1}^n \left(\frac{k_{ij}/nQ}{k_{\cdot j} \cdot 1/n} - 1/\sqrt{n} \right)^2 = \frac{1}{Q} \left(1 - \frac{k_{\cdot j}}{n} \right)$$

Por lo tanto, es aconsejable eliminar las modalidades elegidas muy pocas veces, construyendo otra modalidad uniéndola a la más próxima.

La parte de **inercia debida a una variable** es función creciente del número de modalidades de respuesta que tiene, ya que la inercia de una variable es la suma de las inercias de sus modalidades:

$$I(J_k) = \sum_{j \in J_k} I(j) = \sum_{j \in J_k} \frac{1}{Q} \left(1 - \frac{k_{\cdot j}}{n} \right) = \frac{1}{Q} (m_k - 1)$$

Si una variable tiene un número de modalidades demasiado grande, al igual que en el caso de que su efectivo sea muy pequeño, conviene reagrupar las modalidades en un número que sea razonable y mantener el sentido, para evitar así influencias extremas.

La **inercia total** es la suma de las inercias de todas las modalidades:

$$I = \sum_k I(J_k) = \sum_k \frac{1}{Q} (m_k - 1) = \frac{J}{Q} - 1$$

J/Q es el número medio de modalidades por variable cualitativa o carácter. En consecuencia, la inercia total sólo depende del número de modalidades y del de preguntas.

Si el número de variables es dos, y cada una tiene dos modalidades, los resultados se pueden analizar tanto por Análisis Factorial de Correspondencias (AFC) como por Análisis de Correspondencias Múltiples (ACM). En el primer caso obtendríamos un único factor que recoge el 100% de la inercia total. Esta inercia dependerá del grado de relación que exista entre las modalidades, de modo que, si están poco relacionadas, la inercia será próxima a cero, y si están muy relacionadas, la inercia tenderá a un valor alto.

Si la misma información la analizamos mediante análisis de correspondencias múltiples, obtendremos siempre la misma inercia ($J/Q - 1 = 1$), pero obtendremos dos ejes. En el caso en que exista mucha relación entre las variables, el primer eje recogerá gran parte de la inercia (casi 1) y el segundo muy poca, mientras que en el caso de total independencia entre las dos variables ambos factores recogerán la misma cantidad de inercia, es decir, 1/2 cada uno.

SPSS Y EL ANÁLISIS DE CORRESPONDENCIAS

SPSS Y CORRESPONDENCIAS SIMPLES

SPSS incorpora un procedimiento que implementa el análisis de correspondencias simples. Uno de los fines del análisis de correspondencias es describir las relaciones existentes entre dos variables nominales, recogidas en una tabla de correspondencias, sobre un espacio de pocas dimensiones, mientras que al mismo tiempo se describen las relaciones entre las categorías de cada variable. Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las categorías, con las categorías similares representadas próximas unas a otras. La proyección de los puntos de una variable sobre el vector desde el origen hasta un punto de categoría de la otra variable describe la relación entre ambas variables.

El análisis de las tablas de contingencia a menudo incluye examinar los perfiles de fila y de columna, así como contrastar la independencia a través del estadístico de chi-cuadrado. Sin embargo, el número de perfiles puede ser bastante grande y la prueba de chi-cuadrado no revelará la estructura de la dependencia. El procedimiento Tablas de contingencia ofrece varias medidas y pruebas de asociación pero no puede representar gráficamente ninguna relación entre las variables.

El análisis factorial es una técnica típica para describir las relaciones existentes entre variables en un espacio de pocas dimensiones. Sin embargo, el análisis factorial requiere datos de intervalo y el número de observaciones debe ser cinco veces el número de variables. Por su parte, el análisis de correspondencias asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además, el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas.

Mediante este procedimiento se obtienen medidas de correspondencia, perfiles de fila y de columna, valores propios, puntuaciones de fila y de columna, inercia, masa, estadísticos de confianza para las puntuaciones de fila y de columna, estadísticos de confianza para los valores propios, gráficos de transformación, gráficos de los puntos de fila, gráficos de los puntos de columna y diagramas de dispersión biespaciales.

Para realizar un análisis de correspondencias, elija en los menús *Analizar* → *Reducción de datos* → *Análisis de correspondencias* (Figura 8-1) y seleccione las variables y las especificaciones para el análisis (Figura 8-2). Previamente es necesario cargar en memoria el fichero de nombre COCHES mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene datos sobre automóviles y las variables a analizar son el origen de los coches (*origen*) y su cilindrada (*cilind*).

En cuanto a los datos, las variables categóricas que se van a analizar se encuentran escaladas a nivel nominal. Para los datos agregados o para una medida de correspondencia distinta de las frecuencias, utilice una variable de ponderación con valores de similaridad positivos. De manera alternativa, para datos tabulares, utilice la sintaxis para leer la tabla.

En cuanto a los supuestos, el máximo número de dimensiones utilizado en el procedimiento depende del número de categorías activas de fila y de columna y del número de restricciones de igualdad. Si no se utilizan criterios de igualdad y todas las categorías son activas, la dimensionalidad máxima es igual al número de categorías de la variable con menos categorías menos uno. Por ejemplo, si una variable dispone de cinco categorías y la otra de cuatro, el número máximo de dimensiones es tres. Las categorías suplementarias no son activas. Por ejemplo, si una variable dispone de cinco categorías, dos de las cuales son suplementarias, y la otra variable dispone de cuatro categorías, el número máximo de dimensiones es dos. Considere todos los conjuntos de categorías con restricción de igualdad como una única categoría. Por ejemplo, si una variable dispone de cinco categorías, tres de las cuales tienen restricción de igualdad, dicha variable se debe tratar como si tuviera tres categorías en el momento de calcular la dimensionalidad máxima. Dos de las categorías no tienen restricción y la tercera corresponde a las tres categorías restringidas. Si se especifica un número de dimensiones superior al máximo, se utilizará el valor máximo.

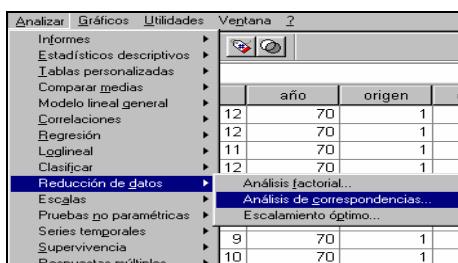


Figura 8-1

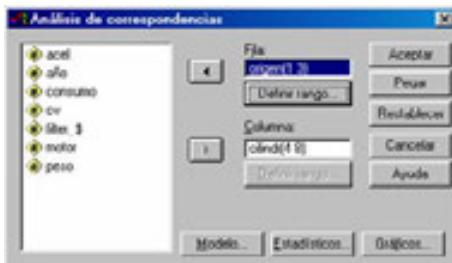


Figura 8-2

En los campos *Fila* y *Columna* de la Figura 8-2 se introducen las dos variables a cruzar en la tabla de contingencia. En los botones *Definir rango* debe definir un rango para las variables de filas (Figura 8-3) y columnas (Figura 8-4). Los valores mínimo y máximo especificados deben ser números enteros. En el análisis, se truncarán los valores de los datos fraccionarios. Se ignorará en el análisis cualquier valor de categoría que esté fuera del rango especificado. Inicialmente, todas las variables estarán sin restringir y activas. Puede restringir las categorías de fila para igualarlas a otras categorías de fila (campo *Restricciones para las categorías*) o puede definir cualquier categoría de fila como suplementaria. *Las categorías deben ser iguales* es una restricción que indica que las puntuaciones de las categorías deben ser iguales. Utilice las restricciones de igualdad si el orden obtenido para las categorías no es el deseado o si no se corresponde con lo intuitivo. El máximo número de categorías de fila que se puede restringir para que sean consideradas iguales es el número total de categorías de fila activas menos 1. Utilice la sintaxis para imponer restricciones de igualdad a diferentes conjuntos de categorías. Por ejemplo, utilice la sintaxis para imponer la restricción de que sean consideradas iguales las categorías 1 y 2 y, por otra parte, que sean consideradas iguales las categorías 3 y 4.

La categoría es suplementaria es una restricción que indica que las categorías suplementarias no influyen en el análisis pero se representan en el espacio definido por las categorías activas. Las categorías suplementarias no juegan ningún papel en la definición de las dimensiones. El número máximo de categorías de fila suplementarias es el número total de categorías de fila menos 2.

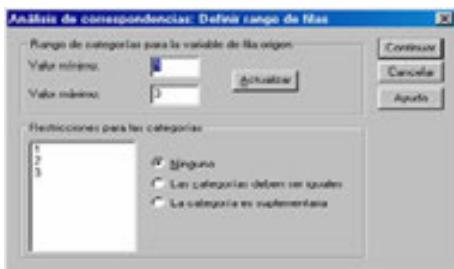


Figura 8-3

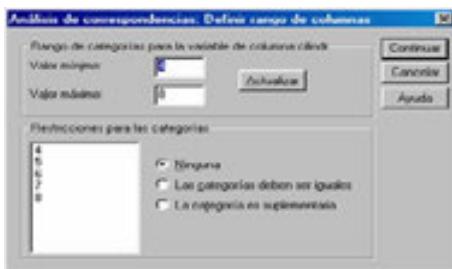


Figura 8-4

El cuadro de diálogo *Modelo* (Figura 8-5) permite especificar el número de dimensiones, la medida de distancia, el método de estandarización y el método de normalización. En la opción *Dimensiones en la solución* especifique el número de dimensiones. En general, seleccione el menor número de dimensiones que necesite para explicar la mayor parte de la variación. El máximo número de dimensiones depende del número de categorías activas utilizadas en el análisis y de las restricciones de igualdad. El máximo número de dimensiones es el menor entre el número de categorías de fila activas menos el número de categorías de fila con restricción de igualdad, más el número de conjuntos de categorías de fila que se han restringido y el número de categorías de columna activas menos el número de categorías de columna con restricción de igualdad, más el número de conjuntos de categorías de columna que se han restringido.

En el cuadro *Medida de distancia* puede seleccionar la medida de distancia entre las filas y columnas de la tabla de correspondencias. Seleccione *Chi-cuadrado* (utiliza una distancia ponderada entre los perfiles, donde la ponderación es la masa de las filas o de las columnas siendo una distancia necesaria para el análisis de correspondencias típico) o *Euclídea* (utiliza la raíz cuadrada de la suma de los cuadrados de las diferencias entre los pares de filas y entre los pares de columnas).

En el cuadro *Método de estandarización* seleccione la opción *Se eliminan las medias de filas y columnas* para centrar las filas y las columnas (este método es necesario para el análisis de correspondencias típico), seleccione *Se eliminan las medias de filas* sólo para centrar las filas, seleccione *Se eliminan las medias de columnas* sólo para centrar las columnas, seleccione *Se igualan los totales de fila y se eliminan las medias* para igualar los márgenes de fila antes de centrar las filas. Seleccione *Se igualan los totales de columna y se eliminan las medias* para igualar los márgenes de columna antes de centrar las columnas.

En el cuadro *Método de normalización* seleccione una de las siguientes opciones:

Simétrico: Para cada dimensión, las puntuaciones de fila son la media ponderada de las puntuaciones de columna divididas por el valor propio coincidente y las puntuaciones de columna son la media ponderada de las puntuaciones de fila divididas por el valor propio coincidente. Utilice este método si desea examinar las diferencias o similaridades entre las categorías de las dos variables.

Principal: Las distancias entre los puntos de fila y los puntos de columna son aproximaciones de las distancias en la tabla de correspondencias de acuerdo con la medida de distancia seleccionada. Utilice este método si desea examinar las diferencias entre las categorías de una o de ambas variables en lugar de las diferencias entre las dos variables.

Principal por fila: Las distancias entre los puntos de fila son aproximaciones de las distancias en la tabla de correspondencias de acuerdo con la medida de distancia seleccionada. Las puntuaciones de fila son la media ponderada de las puntuaciones de columna. Utilice este método si desea examinar las diferencias o similaridades entre las categorías de la variable de filas.

Principal por columna: Las distancias entre los puntos de columna son aproximaciones de las distancias en la tabla de correspondencias de acuerdo con la medida de distancia seleccionada. Las puntuaciones de columna son la media ponderada de las puntuaciones de fila. Utilice este método si desea examinar las diferencias o similaridades entre las categorías de la variable de columnas.

Personalizado: Debe especificar un valor entre -1 y 1. El valor -1 corresponde a *Principal por columna*. El valor 1 corresponde a *Principal por fila*. El valor 0 corresponde a *Simétrico*. Todos los demás valores dispersan la inercia entre las puntuaciones de columna y de fila en diferentes grados. Este método es útil para generar diagramas de dispersión biespaciales a medida.

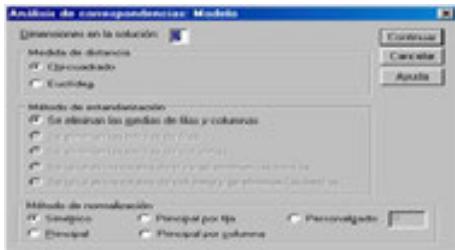


Figura 8-5

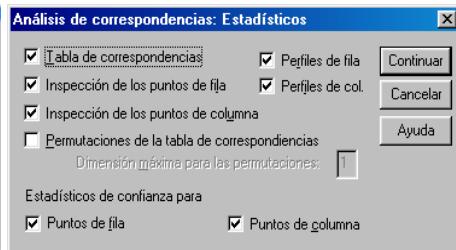


Figura 8-6

El botón *Estadísticos* de la Figura 8-2 nos lleva al cuadro de diálogo *Estadísticos* (Figura 8-6), que permite especificar los resultados numéricos producidos. Las opciones posibles son: *Tabla de correspondencias*, que ofrece la tabla de contingencia de las variables de entrada con los totales marginales de fila y columna; *Inspección de los puntos de fila*, que ofrece para cada categoría de fila las puntuaciones, la masa, la inercia, la contribución a la inercia de la dimensión y la contribución de la dimensión a la inercia del punto; *Inspección de los puntos de columna*, que ofrece para cada categoría de columna las puntuaciones, la masa, la inercia, la contribución a la inercia de la dimensión y la contribución de la dimensión a la inercia del punto; *Perfiles de fila*, que ofrece para cada categoría de fila la distribución a través de las categorías de la variable de columna; *Perfiles de col.*, que ofrece para cada categoría de columna la distribución a través de las categorías de la variable de fila y *Permutaciones de la tabla de correspondencias*, que ofrece la tabla de correspondencias reorganizada de tal manera que las filas y las columnas estén en orden ascendente de acuerdo con las puntuaciones en la primera dimensión.

Si lo desea, puede especificar el número de la dimensión máxima para el que se generarán las tablas permutadas. Se generará una tabla permutada para cada dimensión desde 1 hasta el número especificado. La opción *Estadísticos de confianza para puntos de fila* incluye la desviación típica y las correlaciones para todos los puntos de fila no supplementarios y la opción *Estadísticos de confianza para puntos de columna* incluye la desviación típica y las correlaciones para todos los puntos de columna no supplementarios.

El botón *Gráficos* de la Figura 8-2 nos lleva al cuadro de diálogo *Gráficos* de la Figura 8-7 que permite especificar qué gráficos se van a generar. La opción *Diagramas de dispersión* produce una matriz de todos los gráficos por parejas de las dimensiones.

Los diagramas de dispersión disponibles incluyen: *Diagrama de dispersión biespacial* (produce una matriz de diagramas conjuntos de los puntos de fila y de columna y si está seleccionada la normalización principal, el diagrama de dispersión biespacial no estará disponible), *Puntos de fila* (produce una matriz de diagramas de los puntos de fila), *Puntos de columna* (produce una matriz de diagramas de los puntos de columna). Si lo desea, puede especificar el número de caracteres de etiqueta de valor que se va a utilizar al etiquetar los puntos. Este valor debe ser un entero no negativo menor o igual que 20.

La opción *Gráfico de líneas* produce un gráfico para cada dimensión de la variable seleccionada. Los gráficos de líneas disponibles incluyen: *Categorías de fila transformadas* (produce un gráfico de los valores originales para las categorías de fila frente a las puntuaciones de fila correspondientes) y *Categorías de columna transformadas* (produce un gráfico de los valores originales para las categorías de columna frente a las puntuaciones de columna correspondientes). Si lo desea, puede especificar el número de caracteres de etiqueta de valor que se va a utilizar al etiquetar los ejes de categorías. Este valor debe ser un entero no negativo menor o igual que 20.

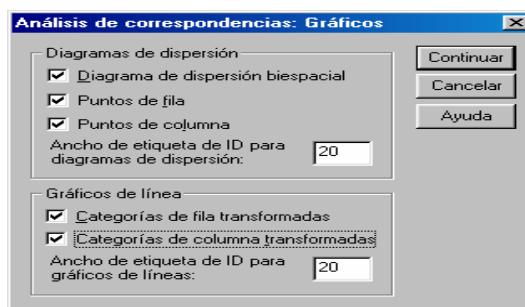


Figura 8-7

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables.

Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 8-2 para obtener los resultados del análisis de correspondencias según se muestra en la Figura 8-8. En la parte izquierda de la figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla.

En las Figuras 8-8 a 8-10 se presentan varias salidas tabulares de entre las múltiples que ofrece el procedimiento.

En las Figuras 8-11 a 8-15 se presentan varias salidas gráficas de entre las múltiples que ofrece el procedimiento.

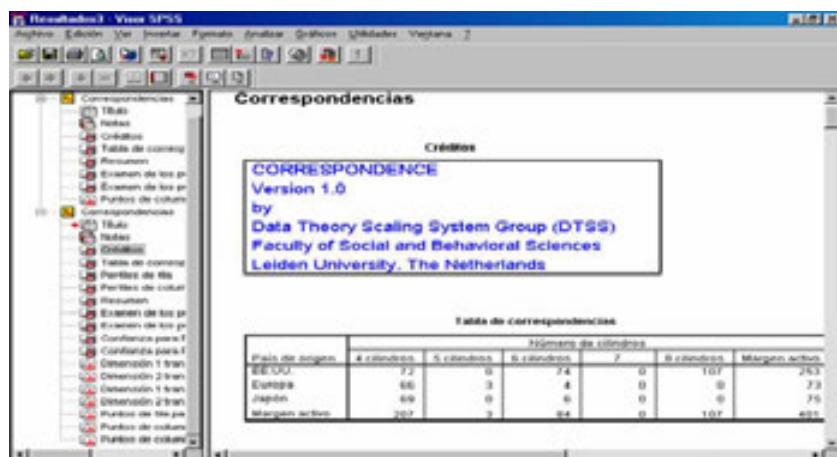


Figura 8-8

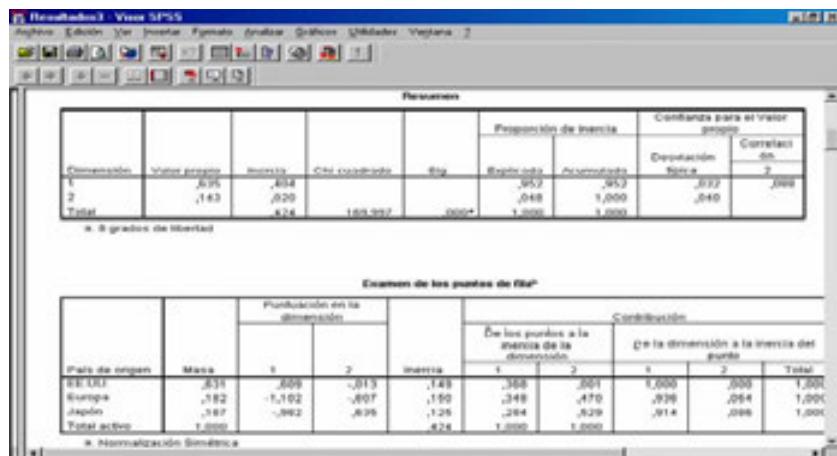


Figura 8-9

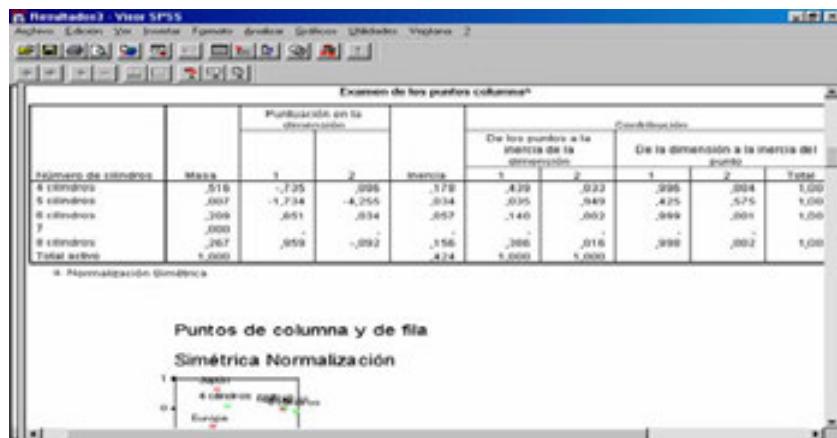


Figura 8-10

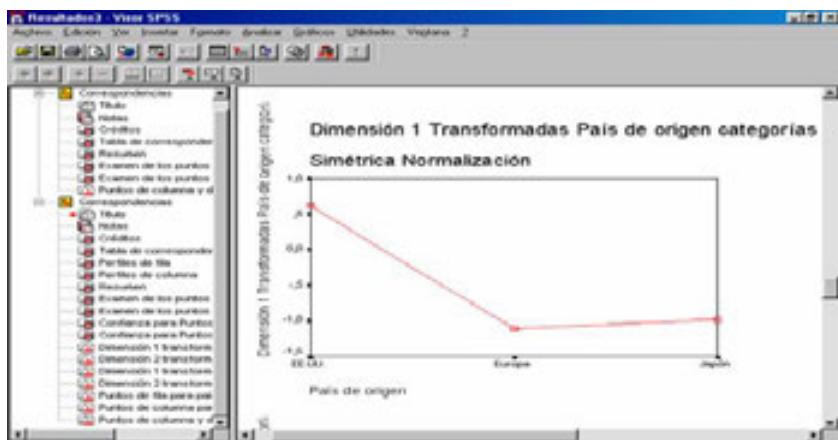


Figura 8-11

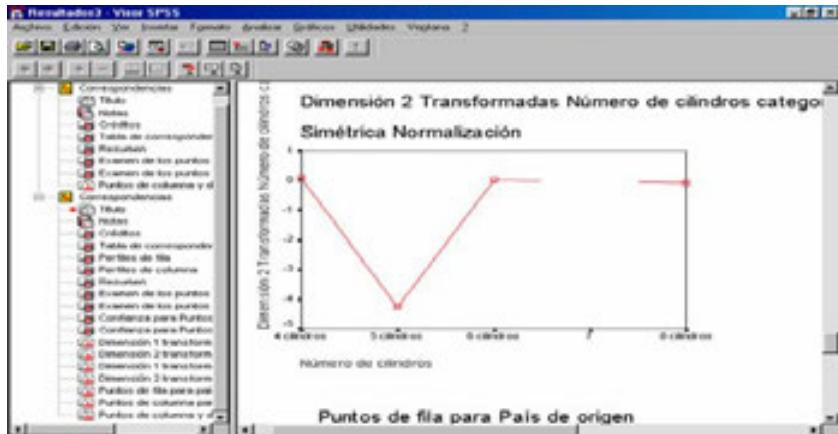


Figura 8-12

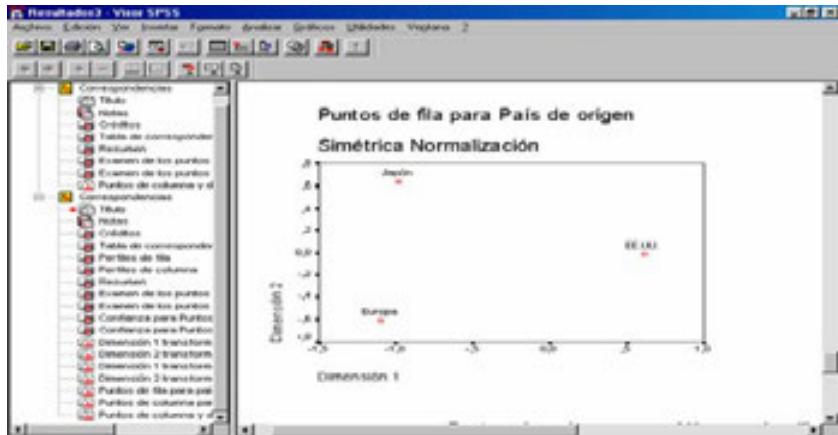


Figura 8-13

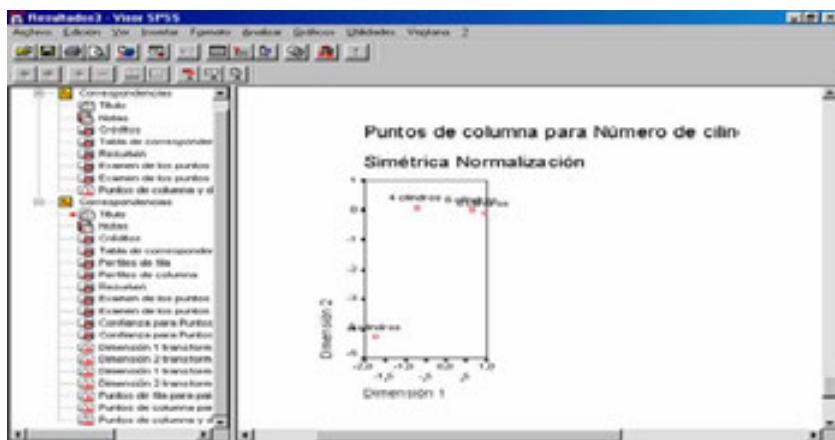


Figura 8-14

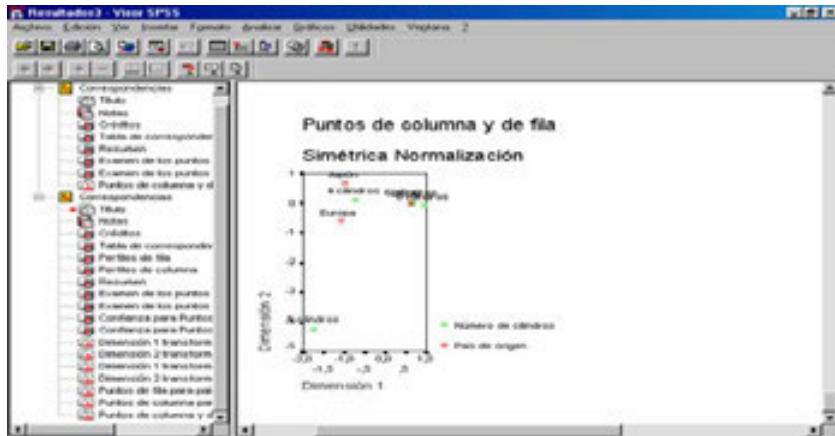


Figura 8-15

La Figura 8-8 muestra la tabla de contingencia para las dos variables con sus marginales. La Figura 8-9 muestra los perfiles de fila y columna, que son las proporciones en cada fila y columna de cada celda basadas en los totales marginales. Los gráficos de puntos fila y columna de las Figuras 8-13 a 8-15 representan estas proporciones para la localización geométrica de los puntos. La Figura 8-10 muestra un cuadro resumen con la solución que representa la relación entre las variables fila y columna en tan pocas dimensiones como es posible. En nuestro caso tenemos dos dimensiones, mostrando la primera una cantidad mayor de inercia (el 95% de la inercia total). Los valores propios pueden interpretarse como la correlación entre las puntuaciones de filas y columnas. Para cada dimensión el cuadrado del valor propio es igual a la inercia y por tanto es otra medida de la importancia de esa dimensión.

En el examen de los puntos fila y columna (Figuras 8-9 y 8-10) se ofrecen las contribuciones a la inercia total de cada punto fila y columna. Los puntos fila y columna que contribuyen sustancialmente a la inercia de una dimensión son importantes para esa dimensión. Los puntos dominantes de la solución pueden detectarse fácilmente. Por ejemplo, Japón es un punto dominante de la segunda dimensión ya que su contribución a la inercia de esa dimensión es 0,635 y Estados Unidos en la primera dimensión pues su contribución es 0,609. Por otra parte, los coches de 8 cilindros (0,959) y 6 cilindros (0,651) contribuyen más que otros a la primera dimensión. A la segunda dimensión los que más contribuyen negativamente son los de 4 cilindros (0,096).

El gráfico de categorías de fila transformadas de la Figura 8-11 produce un gráfico de los valores originales para las categorías de fila frente a las puntuaciones de fila, y el gráfico de categorías de columna transformadas de la Figura 8-12 produce un gráfico de los valores originales para las categorías de columna frente a las puntuaciones de columna.

SPSS Y LAS CORRESPONDENCIAS MÚLTIPLES

SPSS incorpora un procedimiento que implementa el análisis de correspondencias múltiple o análisis de homogeneidades. El análisis de homogeneidad cuantifica los datos (categóricos) nominales mediante la asignación de valores numéricos a los casos (los objetos) y a las categorías. El análisis de homogeneidad se conoce también por el acrónimo HOMALS, del inglés *homogeneity analysis by means of alternating least squares* (análisis de homogeneidad mediante mínimos cuadrados alternantes).

El objetivo de HOMALS es describir las relaciones entre dos o más variables nominales en un espacio de pocas dimensiones que contiene las categorías de las variables así como los objetos pertenecientes a dichas categorías. Los objetos pertenecientes a la misma categoría se representan cerca los unos de los otros, mientras que los objetos de diferentes categorías se representan alejados los unos de los otros. Cada objeto se encuentra lo más cerca posible de los puntos de categoría para las categorías a las que pertenece dicho objeto. El análisis de homogeneidad es similar al análisis de correspondencias simples, pero no está limitado a dos variables. Es por ello que el análisis de homogeneidad se conoce también como el análisis de correspondencias múltiple. También se puede ver el análisis de homogeneidad como un análisis de componentes principales para datos nominales.

El análisis de homogeneidad es más adecuado que el análisis de componentes principales típico cuando puede que no se conserven las relaciones lineales entre las variables, o cuando las variables se miden a nivel nominal. Además, la interpretación del resultado es mucho más sencilla en HOMALS que en otras técnicas categóricas, como pueden ser las tablas de contingencia y los modelos loglineales. Debido a que las categorías de las variables son cuantificadas, se pueden aplicar sobre las cuantificaciones técnicas que requieren datos numéricos, en análisis subsiguientes.

Como ejemplo, el análisis de homogeneidad se puede utilizar para representar gráficamente la relación entre la categoría laboral (*catlab*), la clasificación étnica (*minoría*) y el género (*sexo*) de los empleados de una empresa. Puede que encontremos que la clasificación étnica y el género son capaces de discriminar entre las personas, pero no así la categoría laboral. También puede que encontremos que las categorías *Latino* y *Afro-americano* son similares entre sí. O puede que encontremos cualquier otra relación.

Para realizar un análisis de correspondencias múltiples, elija en los menús *Analizar* → *Reducción de datos* → *Escalamiento óptimo* (Figura 8-16). Previamente es necesario cargar en memoria el fichero de nombre EMPLEADOS mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene datos sobre los trabajadores de una empresa con las variables *catlab*, *minoría* y *sexo* antes descritas.

En el cuadro de diálogo *Escalamiento óptimo* de la Figura 8-17, seleccione *Todas las variables son nominales múltiples*. A continuación seleccione *Un conjunto*, pulse en *Definir*, y en la Figura 8-18 seleccione dos o más variables para el análisis. Defina los rangos para las variables con el botón *Definir rango*. Si lo desea, tiene la posibilidad de seleccionar una o más variables para proporcionar etiquetas de punto en los gráficos de las puntuaciones de objeto (campo *Etiquetar gráficos de las puntuaciones de objeto con*). Cada variable genera un gráfico diferente, con los puntos etiquetados mediante los valores de dicha variable. Debe definir un rango para cada una de las variables de etiquetado de los gráficos. Mediante el cuadro de diálogo, no se puede utilizar una misma variable en el análisis y como variable de etiquetado. Si se desea etiquetar el gráfico de las puntuaciones de objeto con una variable utilizada ya en el análisis, utilice la función *Calcular* en el menú *Transformar* para crear una copia de dicha variable. Utilice la nueva variable para etiquetar el gráfico. Alternativamente, se puede utilizar la sintaxis de comandos. En el botón *Dimensiones en la solución* especifique el número de dimensiones que desea en la solución. En general, seleccione el menor número de dimensiones que necesite para explicar la mayor parte de la variación. Si el análisis incluye más de dos dimensiones, SPSS genera gráficos tridimensionales de las tres primeras dimensiones. Si se edita el gráfico, se pueden representar otras dimensiones.

El botón *Opciones* (Figura 8-19) permite seleccionar estadísticos y gráficos opcionales, guardar en el archivo de datos de trabajo las puntuaciones de los objetos como nuevas variables y, por último, especificar los criterios de iteración y de convergencia. En cuanto a estadísticos y gráficos se obtienen: frecuencias, autovalores, historial de iteraciones, puntuaciones de objeto, cuantificaciones de categoría, medidas de discriminación, gráficos de las puntuaciones de objeto, gráficos de las cuantificaciones de categoría y gráficos de las medidas de discriminación.

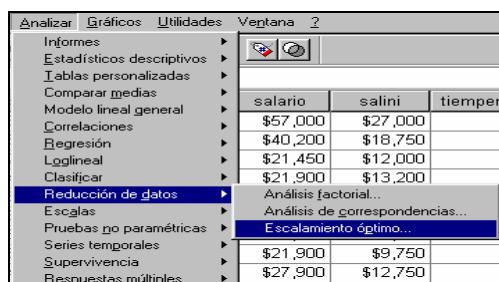


Figura 8-16

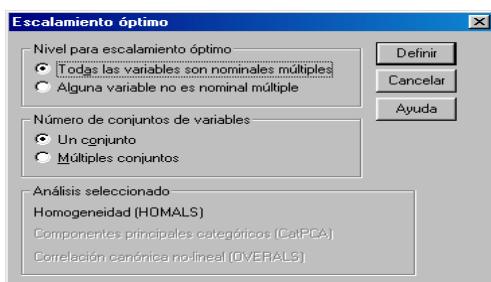


Figura 8-17

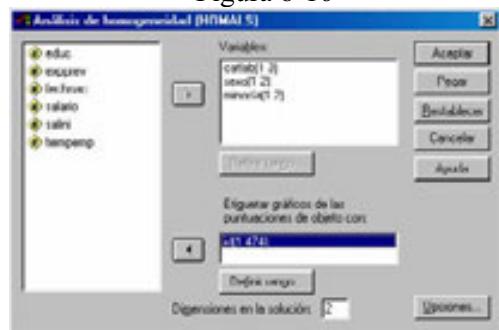


Figura 8-18



Figura 8-19

En cuanto a los datos, todas las variables son nominales múltiples y tienen cuantificaciones de categorías que pueden diferir para cada dimensión. Una vez elegidas las especificaciones (que se aceptan con el botón *Continuar*), se pulsa el botón *Aceptar* en la Figura 8-18 para obtener los resultados del análisis de correspondencias múltiples según se muestra en la Figura 8-20. En la parte izquierda de la figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 8-21 a 8-27 se presentan varias salidas tabulares y gráficas de entre las múltiples que ofrece el procedimiento.

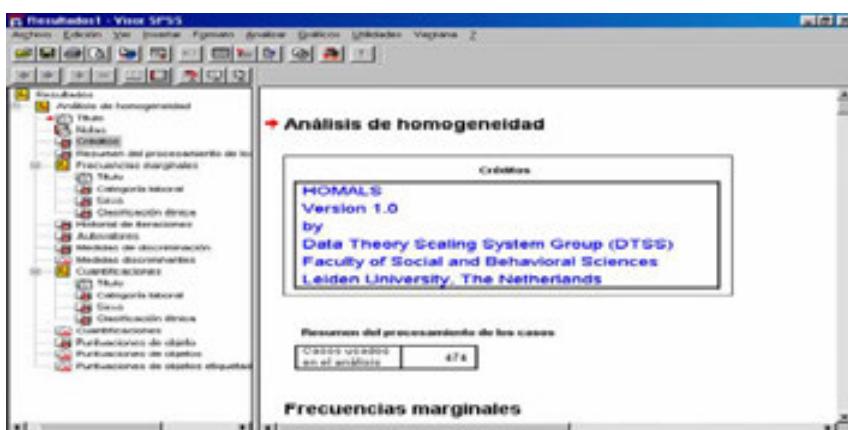


Figura 8-20

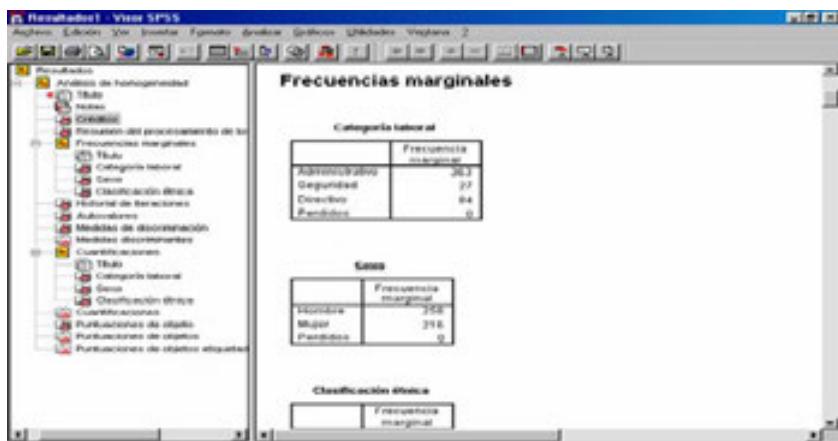


Figura 8-21

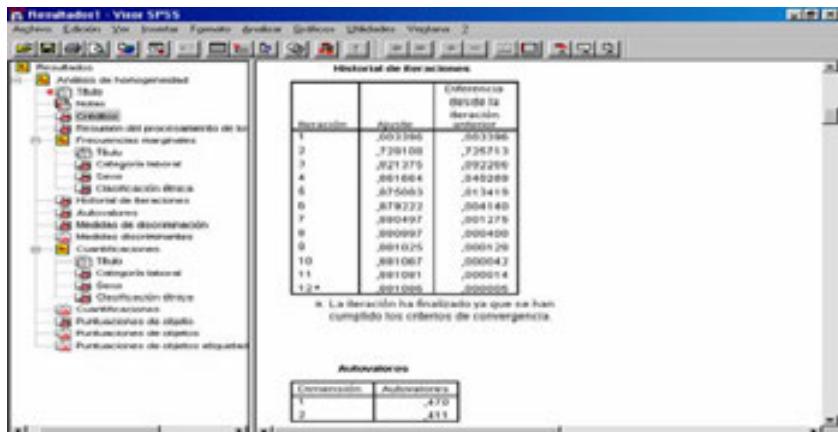


Figura 8-22

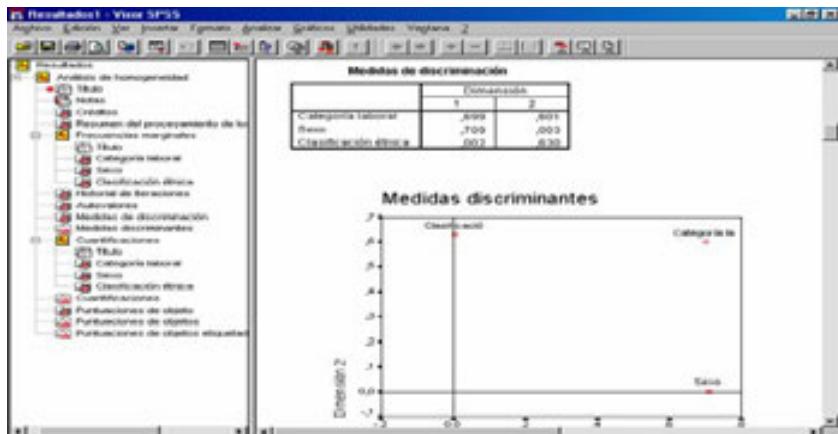


Figura 8-23

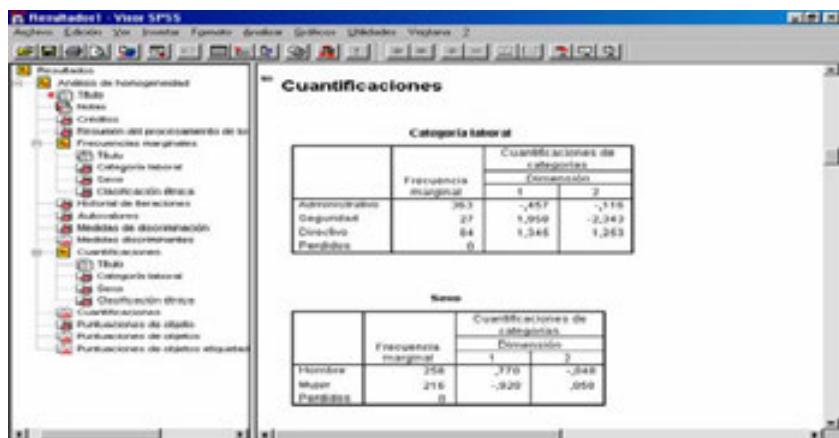


Figura 8-24

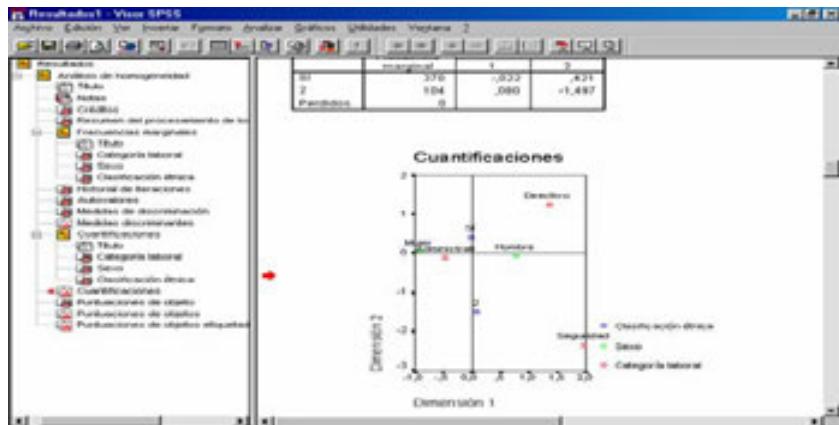


Figura 8-25

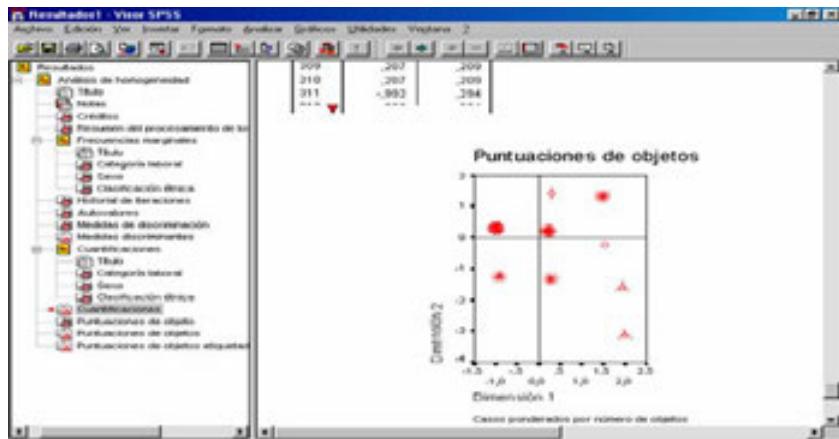


Figura 8-26

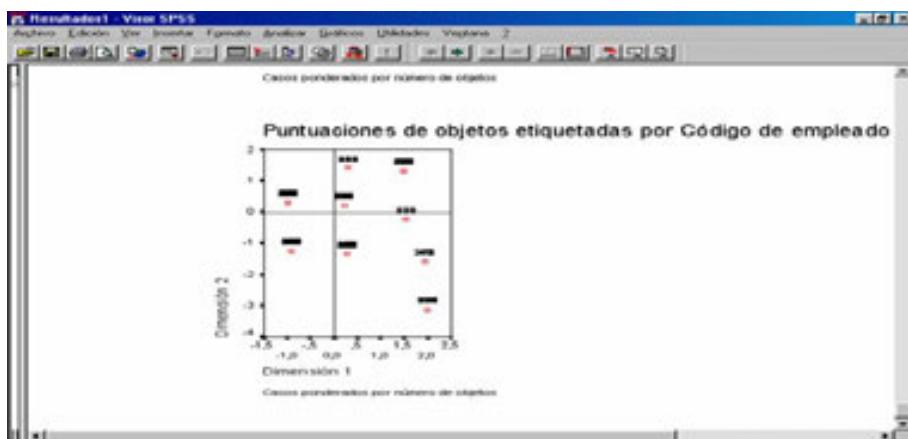


Figura 8-27

En las Figuras 8-20 y 8-21 se muestran resúmenes de casos y tablas de frecuencias marginales representando cada uno de los valores para cada una de las variables. En la Figura 8-22 aparece la historia del proceso de homogeneización a través de las distintas iteraciones que el procedimiento considera necesarias para llegar a una solución de convergencia que refleje el ajuste total, así como la tabla de autovalores para cada dimensión del análisis. Como el análisis se realiza sobre los dos primeros ejes o dimensiones, se muestra en cada una de ellas la medida de la varianza explicada por cada dimensión. La magnitud de esta varianza es una muestra del grado de importancia de dicha dimensión en la solución global. Se observa que las dos dimensiones son casi igual de importantes ya que los dos valores propios están muy próximos.

En la Figura 8-23 también aparecen una serie de medidas de discriminación para cada variable y dimensión, de modo que cuanto más alto sea el valor de la medida de discriminación de una variable determinada en una dimensión dada, más alta será la importancia de dicha variable dentro de esa dimensión. El sexo es más importante en la primera dimensión y la clasificación étnica en la segunda, manteniéndose parecida en las dos dimensiones la influencia de la categoría laboral.

El gráfico de medidas de discriminación de la Figura 8-23 ilustra los resultados de la tabla de medidas de discriminación. De esta manera, la variable *categoría laboral* es la variable líder en el ranking de variables explicativas de la varianza del modelo homogeneizador y las variables menos explicativas son *sexo* y *clasificación étnica*. Por otra parte, medidas de discriminación similares de una variable en todas las dimensiones reflejan dificultades de asignación de la misma a una dimensión dada. Es ideal que una variable tenga un valor alto en una sola dimensión y bajo en la otra (el caso de *sexo* y *clasificación étnica*).

La Figura 8-24 presenta frecuencias marginales y cuantificaciones categóricas para todos los valores de todas las variables. El *gráfico de cuantificaciones* de la Figura 8-25 muestra las cuantificaciones de las categorías etiquetadas con etiquetas de los valores. Las cuantificaciones de las categorías son el promedio de las puntuaciones de los objetos de la misma categoría.

El *gráfico de puntuaciones de objeto* de la Figura 8-26 muestra las citadas puntuaciones, que son medidas representativas de la varianza asignada a cada objeto dentro de cada variable en el contexto de una dimensión particular. Las puntuaciones tienden a valer cero en posiciones de equilibrio, es decir, cuando el objeto no ejerce ningún papel claro en ninguna dirección. A medida que el valor es más alto hay mayor tendencia en el objeto a estar representado por el análisis de homogeneidades realizado siguiendo sus pautas.

El *gráfico de puntuaciones de los objetos etiquetados* de la Figura 8-27 es muy útil para mostrar objetos que constituyen valores atípicos. Por otra parte, en este gráfico se observa si caen juntos muchos objetos. El eje horizontal de este gráfico se corresponde con la primera dimensión y el gráfico sirve para ver si por encima y por debajo del eje horizontal hay alguna agrupación de objetos homogéneos tales que el eje los discrimine bien. El eje vertical del gráfico se corresponde con la segunda dimensión y el gráfico sirve para ver si a la izquierda y a la derecha del eje vertical hay alguna agrupación de objetos homogéneos tales que el eje los discrimine bien.

APLICACIONES DEL ANÁLISIS DE CORRESPONDENCIAS

Ya sabemos que el análisis de correspondencias es un método de interdependencia cuyo objetivo es la reducción de la dimensión en el sentido de transformar la información de un conjunto elevado de variables categóricas a un conjunto menor que las represente convenientemente. Esta reducción de la dimensión se lleva a cabo profundizando en las relaciones que se establecen entre los valores o categorías de las distintas variables. El descubrimiento de estas relaciones se apoya bastante en las representaciones gráficas de la información de las tablas de contingencia (que concentran la información de las variables cualitativas en estudio) para visualizar la proximidad o lejanía entre las categorías que forman parte de las variables.

Cuando se aplica en la práctica un método de análisis factorial con variables cuantitativas hay que tener presente que el número de observaciones ha de ser por lo menos cinco veces superior al de variables y además sólo será posible detectar relaciones lineales entre las variables.

Sin embargo, cuando se aplica en la práctica un método factorial con variables cualitativas, se puede detectar cualquier tipo de relación entre las variables, aunque sea no lineal. De hecho, la presencia de relaciones no lineales en las variables es muy habitual en materias como la investigación comercial al considerar los beneficios marginales.

Por otra parte, los métodos factoriales con variables categóricas permiten describir las relaciones entre las categorías de las variables y entre las propias variables. Además, las variables categóricas son de mayor uso en la práctica y cualquier variable puede transformarse a categórica.

El análisis de correspondencias se puede aplicar a grandes tablas de contingencia cuyas filas representan los sujetos del análisis, cuyas columnas pueden representar diversos aspectos medidos y cuyo interior expresa el número de individuos que asocian las modalidades de las variables fila con las modalidades de las variables columna. Este número puede ser absoluto, relativo, una proporción etc. También se aplica el análisis de correspondencias a tablas de contingencia que cruzan dos o más variables nominales, a tablas de magnitudes homogéneas donde se describen las características de diversas poblaciones, a tablas de datos ordinales categorizados (escalas de actitud, valoración de marcas, etc.), a tablas de un conjunto heterogéneo de variables codificadas en forma disyuntiva completa (tablas de Burtz), a tablas de datos binarios (por ejemplo, que reflejan presencia o ausencia), a tablas de proximidad o distancia entre elementos, a tablas de correlación, a tablas múltiples de números positivos con tres o más entradas (marcas, atributos, estilos de vida, etc.), a tablas de correlación, a tablas mixtas de variables cualitativas y cuantitativas, a tablas yuxtapuestas, a tablas de valoración o intensidad que recogen preferencias mediante puntuaciones numéricas en vez de frecuencias y, en general, puede utilizarse el análisis de correspondencias en tablas que contengan cualquier medida de correspondencia entre filas y columnas, y referidas a su similitud, afinidad, confusión, asociación, interacción, distancia, etc. Lógicamente será exigible en cualquier análisis de correspondencias que se pueda dar algún sentido a la suma de los casos por filas o columnas en la tabla de entrada de datos para el análisis, es decir, será exigible que la tabla analizada sea una tabla de contingencia.

De lo expuesto en el párrafo anterior se deduce la fuerte aplicación del análisis de correspondencias en el marketing (estudios sobre imagen y posicionamiento, elección de marcas en función del tipo de producto, identificación de claves para la comunicación, detección de la imagen de los elementos que forman parte de un tipo de producto, clasificación y estructura de mercados, contenido de los soportes de medios de comunicación y segmentación de mercados).

En el párrafo anterior solamente se ha hablado de aplicaciones en análisis estáticos, pero el análisis de correspondencias también permite realizar estudios dinámicos para analizar la evolución entre dos momentos del tiempo, como es el caso del cambio en el posicionamiento de marcas a lo largo del tiempo por los efectos de la publicidad o por acciones específicas de marketing. Es conveniente no olvidar que el análisis de correspondencias, inicialmente pensado para aplicaciones de tablas de contingencia, se aplica en la actualidad a cualquier tabla de números positivos, ya sean variables nominales, ordinales, disyuntivas, series temporales, etc.

Ejercicio 8-1 Consideramos la tabla que muestra la distribución hipotética de los asientos del parlamento europeo entre los partidos políticos de 5 naciones:

	<i>Dem.crist.</i>	<i>Socialista</i>	<i>Otros</i>	<i>TOTAL</i>
Bélgica	8	9	7	24
Alemania	39	30	6	75
Italia	25	11	39	75
Luxemburgo	3	2	1	6
Holanda	13	10	2	25
<i>TOTAL</i>	88	62	55	205

Los totales marginales de fila muestran que los países más pequeños tienen menor representación en el parlamento europeo que los países más grandes. Los totales de columna indican que los demócratas cristianos se separan de los socialistas y ambos se separan de los otros. Pero ¿qué tienen de común los países en relación con la afiliación política? y ¿cuál es la relación entre país y partido político?

Cargamos el fichero 8-1sav y usamos un análisis de correspondencias simples rellenando la pantalla de entrada del procedimiento *Análisis de correspondencias* como se indica en la Figura 8-28. Las pantallas de los botones *Modelo*, *Estadísticos* y *Gráficos* se presentan en las Figuras 8-29 8-31 La salida se ve en las Figuras 8-32 8-39

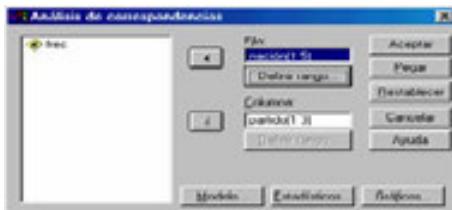


Figura 8-28



Figura 8-29

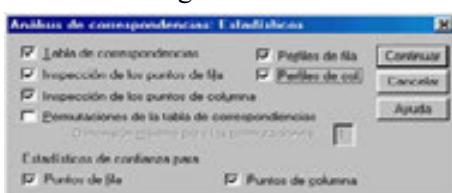


Figura 8-30

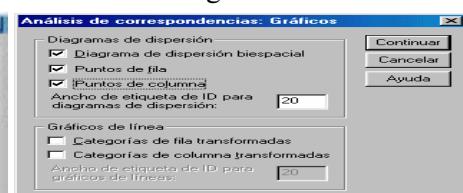


Figura 8-31

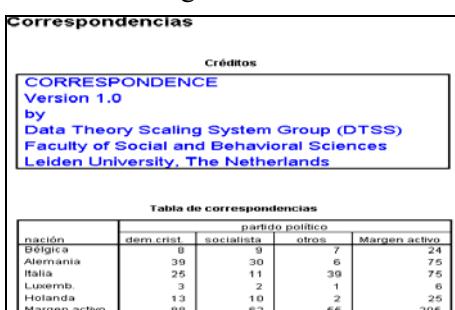


Figura 8-32

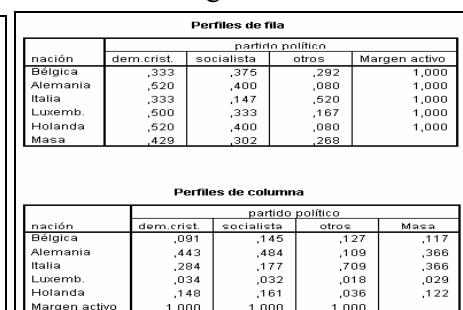


Figura 8-33

Resumen									
Dimensión	Valor propio	Inercia	Chi-cuadrado	Sig.	Proporción de inercia		Confianza para el Valor propio		Correlación
					Explicada	Acumulada	Desviación típica	2	
1	,462	,214			,975	,975	,060		
2	,074	,005			,025	1,000	,069		
Total		,219	44,917	,0000 ^a	1,000	1,000			

a. 8 grados de libertad

Examen de los puntos de fila^a

nación	Masa	Puntuación en la dimensión		Inercia	Contribución				Total	
		1	2		De los puntos a la inercia de la dimensión		De la dimensión a la inercia del punto			
		1	2		1	2	1	2		
Bélgica	,117	-,029	,742	,005	,000	,876	,010	,990	1,000	
Alemania	,366	,628	-,088	,067	,312	,038	,997	,003	1,000	
Italia	,366	-,854	-,101	,124	,577	,051	,998	,002	1,000	
Luxemb.	,029	,326	-,235	,002	,007	,022	,924	,076	1,000	
Holanda	,122	,628	-,088	,022	,104	,013	,997	,003	1,000	
Total activo	1,000		,219	1,000	1,000					

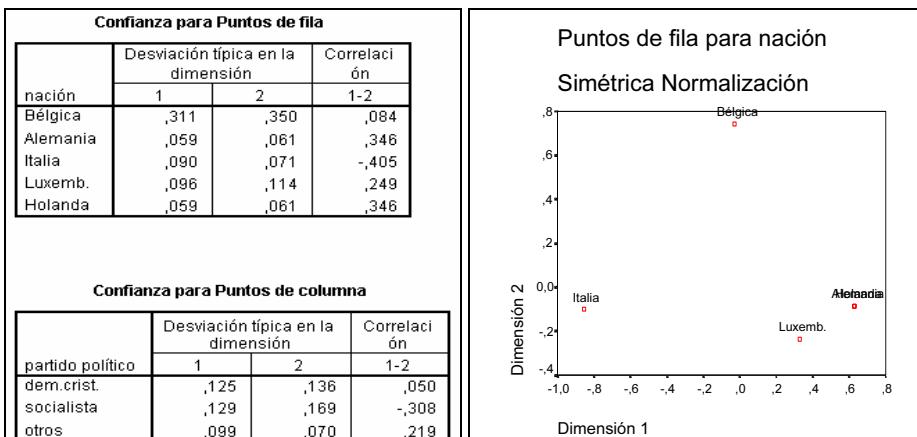
a. Normalización Simétrica

Figura 8-34

Examen de los puntos columna^a

partido político	Masa	Puntuación en la dimensión		Inercia	Contribución				Total		
		De los puntos a la inercia de la dimensión			De la dimensión a la inercia del punto						
		1	2		1	2	1	2			
dem.crist.	,429	,296	-,290	,020	,081	,489	,868	,132	1,000		
socialista	,302	,562	,346	,047	,207	,491	,943	,057	1,000		
otros	,268	-1,107	,074	,152	,712	,020	,999	,001	1,000		
Total activo	1,000			,219	1,000	1,000					

Figura 8-35



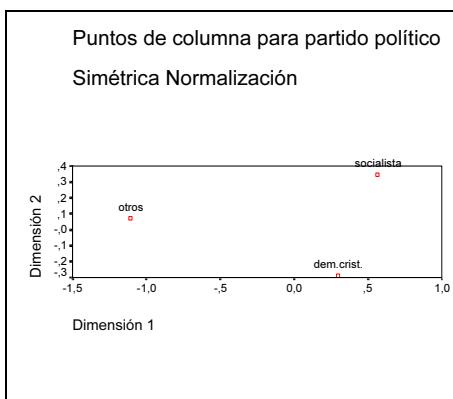


Figura 8-38

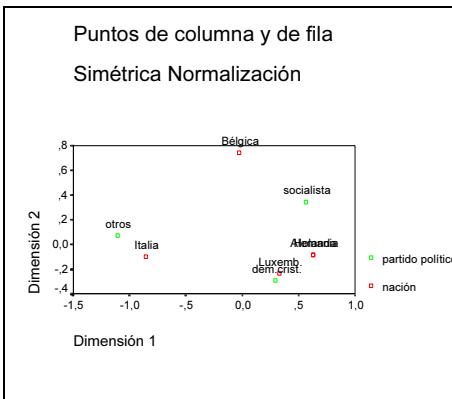


Figura 8-39

Para determinar la distancia entre categorías, el procedimiento considera las distribuciones marginales, así como las frecuencias individuales de cada celda, calculando los perfiles de las filas y las columnas, que son las proporciones en cada fila y columna de cada celda basadas en los totales marginales. En los perfiles de fila de la Figura 8-33 se muestra que los porcentajes de fila para Alemania y Holanda son los mismos, razón por la cual el gráfico de Puntos de fila de la Figura 8-37 coloca a estos dos países en la misma localización.

El cuadro *Resumen* de la Figura 8-34 muestra cómo las filas de una masa pequeña influyen en la inercia sólo cuando están lejos del centroide. Las filas con una gran masa, como Alemania e Italia, influyen en la inercia total sólo cuando están localizadas cerca del centroide. Lo mismo se aplica a las columnas. Cuando las entradas de correspondencias son frecuencias, la suma ponderada de todas las distancias cuadradas entre los perfiles de fila y el perfil medio de fila es igual al estadístico chi-cuadrado. Las distancias euclídeas de la Figura 8-37 aproximan las distancias chi-cuadrado de la tabla. La inercia total se define como la suma ponderada de todas las distancias al centroide dividido por la suma de todas las celdas de la tabla.

En el análisis de correspondencias se busca una solución que represente la relación entre las variables fila y columna en tan pocas dimensiones como sea posible. En nuestro caso (Figura 8-34) tenemos dos dimensiones, mostrando la primera una cantidad mayor de inercia (el 98% de la inercia total). Los valores propios pueden interpretarse como la correlación entre las puntuaciones de filas y columnas. Para cada dimensión el cuadrado del valor propio es igual a la inercia y por tanto es otra medida de la importancia de esa dimensión.

Las puntuaciones de filas de la Figura 8-34 son las coordenadas de los puntos de las filas de la Figura 8-37 y las de las columnas son las coordenadas de los puntos de las columnas de la Figura 8-38. Geométricamente, los puntos de las columnas son proporcionales al centroide ponderado de los puntos de las filas. Las distancias euclídeas entre los puntos de fila aproximan las distancias chi-cuadrado.

En el examen de los puntos fila y columna de las Figuras 8-34 y 8-35 se ofrecen las contribuciones a la inercia total de cada punto fila y columna. Los puntos fila y columna que contribuyen sustancialmente a la inercia de una dimensión son importantes para esa dimensión. Los puntos dominantes de la solución pueden detectarse fácilmente. Por ejemplo, Bélgica es un punto dominante de la segunda dimensión ya que su contribución a la inercia de esa dimensión es 0,742 y Alemania en la primera dimensión pues su contribución es 0,628. Por otra parte, demócratas-cristianos (0,296) y socialistas (0,562) contribuyen más que otros a la primera dimensión. A la segunda dimensión los que más contribuyen son los socialistas (0,346).

En la Figura 8-36 se observa que las desviaciones típicas en filas y columnas para las dos dimensiones son pequeñas, por lo que se puede concluir que la solución obtenida en el análisis de correspondencias es estable. Además si nos fijamos en las correlaciones entre las dimensiones para las puntuaciones también resultan pequeñas.

El gráfico de puntos de fila y columna de la Figura 8-39 muestra que Italia es el país más cercano a otros, Luxemburgo es el más cercano a democristianos y Bélgica, Alemania y Holanda son los más cercanos a socialistas. El gráfico de la Figura 8-37 muestra que Alemania y Holanda son virtualmente idénticas en sus elecciones de partidos políticos y Luxemburgo está muy cerca de ellos, mientras que Italia y Bélgica se encuentran más alejados de los anteriores y también entre sí. En cuanto a partidos políticos, en la Figura 8-38 se ve que socialistas y democristianos están más cerca entre sí que del grupo otros.

Ejercicio 8-2. En el fichero 8-2.sav se almacena un conjunto de datos recolectados con la finalidad de estudiar la realidad socioeconómica española. En este fichero se encuentran 4 variables categóricas con la misma estructura que reflejan en cada uno de los 120 individuos de la muestra los distintos niveles de preocupación y satisfacción suscitados respectivamente por el afecto, el dinero y las relaciones. También se ha codificado la situación o no de empleo de los individuos, pero sólo a efectos de etiquetado. Ante estos datos se trata de mostar y medir la relación de semejanza mutua entre las categorías de las cuatro variables, suponiendo que las respuestas de los individuos encuestados en las cuatro variables analizadas se asocian entre sí.

Comenzamos rellenando la pantalla de entrada del procedimiento *Escalamiento óptimo* como se indica en la Figura 8-40. Se pulsa en *Definir* y se rellena la pantalla de *Análisis de homogeneidades* como se indica en la Figura 8-41. La pantalla del botón *Opciones* se rellena como se indica en la Figura 8-42. Al pulsar *Continuar* y *Aceptar* se obtiene la salida de las Figuras 8-43 a 8-50

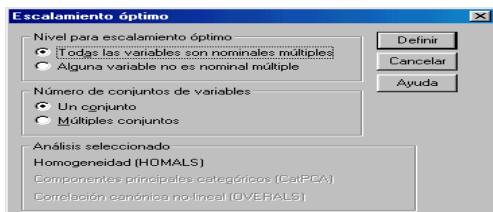


Figura 8-40

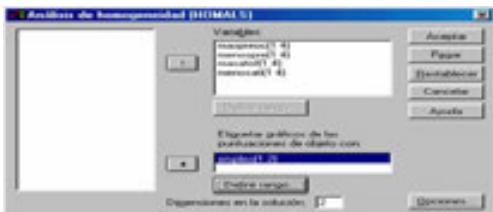


Figura 8-41

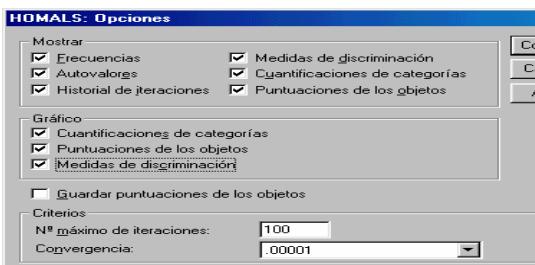


Figura 8-42

Frecuencias marginales	
+preoc	
afecto	Frecuencia marginal
dinero	13
relac	16
dinero	27
Perdidos	59
preoc	
afecto	Frecuencia marginal
dinero	23
relac	53
dinero	13
Perdidos	18

Figura 8-43

+satisf	
	Frecuencia marginal
afecto	45
dinero	42
relac	23
dinero	5
Perdidos	5

-satisf	
	Frecuencia marginal
afecto	15
dinero	5
relac	25
dinero	67
Perdidos	8

Figura 8-44

Historial de iteraciones		
Iteración	Ajuste	Diferencia desde la iteración anterior
1	,041410	,041410
2	,665124	,623715
3	,887993	,022869
4	,703736	,015742
5	,717751	,014015
6	,732415	,014664
7	,748907	,016492
8	,767701	,018794
9	,798647	,020945
10	,810830	,022184
11	,832590	,021759
12	,852028	,019439
13	,867829	,015801

Figura 8-45

Autovalores	
Dimensión	Autovalores
1	,470
2	,435

Medidas de discriminación		
	Dimensión	
	1	2
+preoc	,584	,444
-preoc	,602	,737
+satisf	,354	,294
-satisf	,338	,264

Figura 8-46

CUANTIFICACIONES		
+preoc		
	Frecuencia marginal	Cuantificaciones de categorías
		Dimensión
		1 2
afecto	13	-1,277 -,195
dinero	16	-1,344 1,837
relac	27	,755 ,082
dinero	59	,278 -,406
Perdidos	5	

-preoc		
	Frecuencia marginal	Cuantificaciones de categorías
		Dimensión
		1 2
afecto	23	,610 ,057
dinero	53	,347 ,306
relac	13	-1,563 -1,461
dinero	13	-1,211 2,070
Perdidos	18	

Figura 8-47

+satisf		
	Frecuencia marginal	Cuantificaciones de categorías
		Dimensión
		1 2
afecto	45	-,669 -,142
dinero	42	,545 ,-,014
relac	23	,435 ,873
dinero	5	-1,049 -1,837
Perdidos	5	

-satisf		
	Frecuencia marginal	Cuantificaciones de categorías
		Dimensión
		1 2
afecto	15	1,290 ,821
dinero	5	1,150 ,-,207
relac	25	-,567 -,933
dinero	67	,121 ,242
Perdidos	8	

Figura 8-48

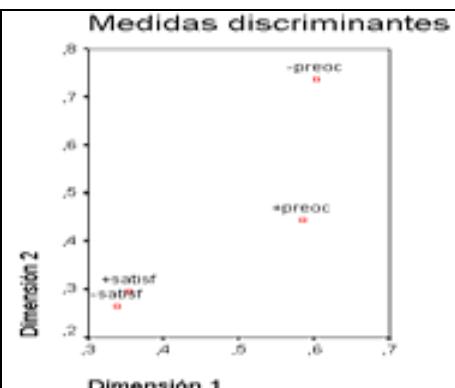


Figura 8-49

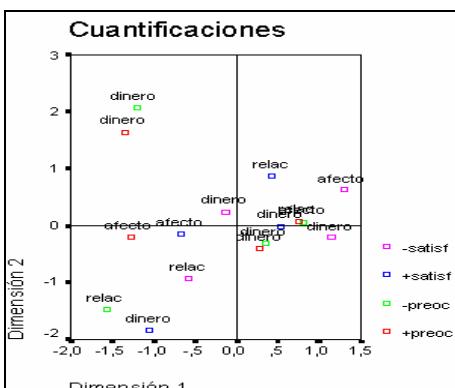


Figura 8-50

Puntuaciones de objeto		
	Dimensión	
	1	2
1	1,675	,459
2	-,518	-,432
3	-,325	-1,028
4	,321	-,953
5	-,815	-,234
6	,746	,441
7	-,733	2,109
8	-,325	-1,028
9	-1,422	-,328
10	-,325	-1,028
11	,808	-,066
12	,734	1,586
13	-,325	-1,028
14	-,815	-,234
15	-,087	-,352

Figura 8-51

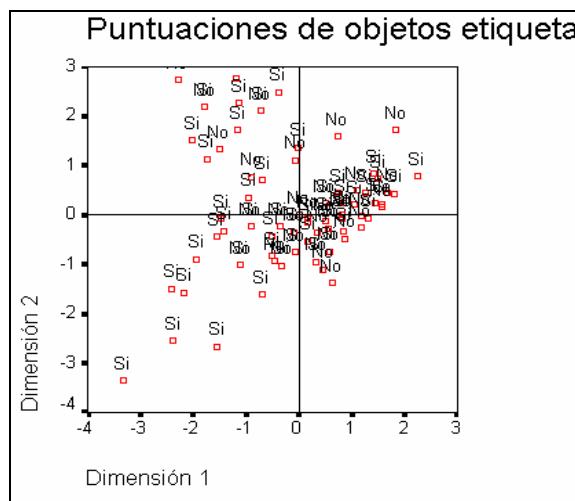


Figura 8-52

En las Figuras 8-43 y 8-44 se muestran tablas de frecuencias marginales representando cada uno de los cuatro valores para cada una de las cuatro variables. En la Figura 8-45 aparece la historia del proceso de homogeneización a través de las distintas iteraciones que el procedimiento considera necesarias para llegar a una solución de convergencia que refleje el ajuste total.

En la Figura 8-46 aparece la tabla de autovalores para cada dimensión del análisis. Como el análisis se realiza sobre los dos primeros ejes o dimensiones, se muestra en cada una de ellas la medida de la varianza explicada por cada dimensión. La magnitud de esta varianza es una muestra del grado de importancia de dicha dimensión en la solución global. Se observa que las dos dimensiones son casi igual de importantes ya que los dos valores propios son muy próximos. En la Figura 8-46 también aparecen una serie de medidas de discriminación para cada variable y dimensión, de modo que cuanto más alto sea el valor de la medida de discriminación de una variable determinada en una dimensión dada, más alta será la importancia de dicha variable dentro de esa dimensión. De esta manera, la variable *-preoc* es la variable líder en el ranking de variables explicativas de la varianza del modelo homogeneizador (ver el **gráfico de medidas de discriminación** de la Figura 8-49 en el que se refleja que las variables menos explicativas son *menossatif* y *masatif*, seguidas de *maspreoc*). Además, esta variable es la que más varianza explicada condensa. Por otra parte, medidas de discriminación similares de una variable en todas las dimensiones reflejan dificultades de asignación de la misma a una dimensión dada. Es ideal que una variable tenga un valor alto en una sola dimensión y bajo en la otra.

Las Figuras 8-47 y 8-48 presentan frecuencias marginales y cuantificaciones categóricas para todos los valores de todas las variables. El *gráfico de cuantificaciones* de la Figura 8-50 muestra las cuantificaciones de las categorías etiquetadas con etiquetas de los valores. Las cuantificaciones de las categorías son el promedio de las puntuaciones de los objetos de la misma categoría.

Las puntuaciones de objeto de la Figura 8-51 son medidas representativas de la varianza asignada a cada objeto dentro de cada variable en el contexto de una dimensión particular. Las puntuaciones tienden a valer cero en posiciones de equilibrio, es decir, cuando el objeto no ejerce ningún papel claro en ninguna dirección. A medida que el valor es más alto hay mayor tendencia en el objeto a estar representado por el análisis de homogeneidades realizado siguiendo sus pautas. La Figura 8-52 muestra el *gráfico de puntuaciones de los objetos* etiquetados que es muy útil para mostrar objetos que constituyen valores atípicos. Por otra parte, en este gráfico se observa si caen juntos muchos objetos. El eje horizontal de este gráfico se corresponde con al primer dimensión y el gráfico sirve para ver si por encima y por debajo del eje horizontal hay alguna agrupación de objetos homogéneos tales que el eje los discrimine bien. El eje vertical del gráfico se corresponde con la segunda dimensión y el gráfico sirve para ver si a la izquierda y a la derecha del eje vertical hay alguna agrupación de objetos homogéneos tales que el eje los discrimine bien.

ESCALAMIENTO ÓPTIMO Y MULTIDIMENSIONAL

CONCEPTO DE ESCALAMIENTO ÓPTIMO

Las diversas técnicas de escalamiento óptimo se caracterizan porque manejan datos categóricos. Los datos categóricos se utilizan a menudo en investigación de mercados, investigación de informes, e investigación en Ciencias Sociales. De hecho, muchos investigadores trabajan exclusivamente con datos categóricos. Los análisis categóricos se resumen típicamente en tablas de contingencia.

En el caso de variables categóricas el análisis de datos tabulares requiere un conjunto de modelos estadísticos diferentes de las típicas aproximaciones correlación y regresión, usadas para datos no categóricos, es decir, cuantitativos. El análisis tradicional de las tablas de dos dimensiones consiste en mostrar frecuencias de celdas, además de uno o más porcentajes de esa celda. Si los datos de la tabla representan una muestra, se podría calcular el estadístico chi-cuadrado, junto con una o más medidas de asociación.

Las tablas de muchas dimensiones se manejan con dificultad, ya que la visualización de los datos está influida por el conocimiento de cuál es la variable fila, cuál la variable columna y cuál la(s) variable(s) de control. Los métodos tradicionales de tablas de contingencia no funcionan bien para tres o más variables, porque todos los estadísticos a producir son estadísticos condicionales, que no retratan, en general, las relaciones entre variables.

Se han desarrollado varios tipos de *modelos loglineales* como un medio comprensible de afrontar las tablas de dos y más dimensiones que se estudiarán en capítulos posteriores. Con la misma finalidad y alternativamente aparecen los *modelos de escalamiento óptimo*.

En cuanto a las ventajas de los modelos loglineales contamos con que hay muchos modelos comprensibles para aplicar a tablas con complejidad arbitraria. Por otra parte, los modelos loglineales proporcionan estadísticos de bondad del ajuste y permiten acometer la construcción del modelo hasta que se encuentra un modelo adecuado ofreciendo estimaciones de los parámetros y errores estándar

En cuanto a las desventajas de los modelos loglineales, si el tamaño de la muestra es demasiado pequeño, se deberá sospechar del estadístico chi-cuadrado en el que se basan los modelos. Si el tamaño de la muestra es demasiado grande, es difícil llegar a un modelo parsimonioso, y puede resultar difícil discriminar entre los modelos “competidores” que aparecen para ajustar los datos. Según el número de variables y el de valores por variable aumentan, se necesitan modelos con más parámetros, y surgen dificultades en interpretar las estimaciones de los parámetros. Es precisamente en estas situaciones de desventaja cuando la técnicas de escalamiento óptimo toman su mejor dimensión.

Suelen utilizarse cuatro procedimientos relacionados con la ejecución del *Escalamiento Óptimo* que son los siguientes:

Análisis de Correspondencias Simples (ANACOR): Analiza tablas de contingencia de 2 dimensiones y ya fue analizado en un capítulo anterior.

Análisis de Correspondencias Múltiples u Homogeneidades (HOMALS): Analiza datos de tabla de contingencia de múltiples dimensiones, donde todas las variables utilizadas son de nivel nominal y donde pueden ignorarse las interacciones de más dimensiones. También ha sido ya tratado en un Capítulo anterior

Análisis de Componentes Principales Categóricas (CATPCA): Contabiliza los patrones de variación en un sólo conjunto de variables de niveles de medición mixtos.

Análisis No Lineal de Correlación Canónica (OVERALS): Corrobora la extensión a la que se correlacionan 2 o más conjuntos de variables de niveles de medición mixtos.

Estos procedimientos son técnicas de Reducción de datos (dimensiones), que intenta representar las múltiples relaciones entre variables en un número de dimensiones reducido. Esto permite describir estructuras o patrones en las relaciones entre variables, difícilmente observables de otro modo. Estas técnicas pueden ser una forma de representación cartográfica perceptual (*perceptual mapping*). Una gran ventaja de estos procedimientos es que acomodan los datos a los diferentes niveles de medida.

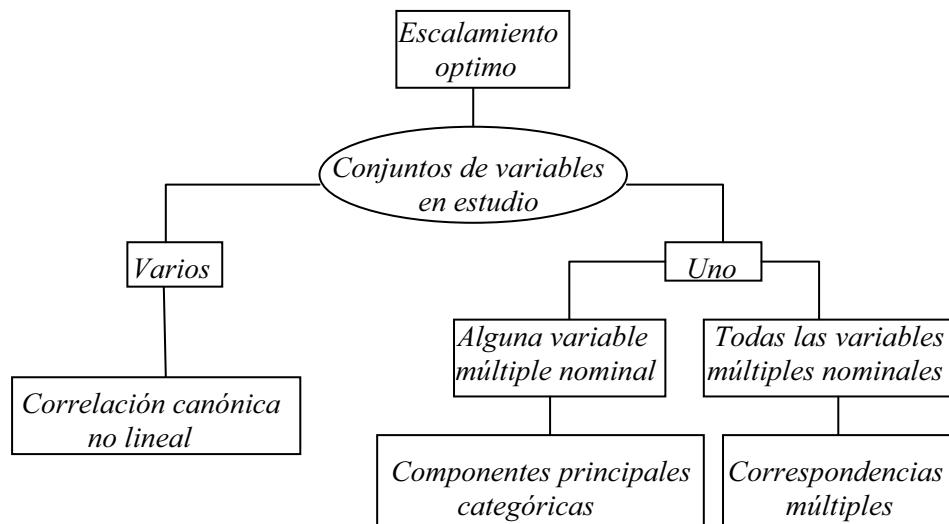
En el análisis estadístico estándar, el nivel de medida es una propiedad fija de cada variable. El nivel de medida orienta la elección de una técnica apropiada. En el escalamiento óptimo, el nivel de medida se contempla normalmente como una opción especificada por el usuario. Ajustando el nivel especificado de medida de algunas variables en el análisis, se podrían descubrir relaciones ocultas. En el escalamiento óptimo existen los siguientes niveles de medida:

Nivel nominal: Los valores de variable representan categorías no ordenadas. Ejemplos de variables con nivel nominal de medida son región, provincia o afiliación religiosa.

Nivel ordinal: Los valores de variable representan valores ordenados. Ejemplos: escalas de actitud (grado de satisfacción o confianza) o las puntuaciones de prorratoe de preferencia.

Nivel numérico (de intervalo): Los valores de variable representan categorías ordenadas, con una métrica significativa, de forma que tiene sentido comparar la distancia entre categorías. Ejemplos: la edad en años o ingresos en miles de dólares.

Podríamos representar los procedimientos de escalamiento óptimo como sigue:



El que una variable sea intrínsecamente numérica no significa que una relación con otra variable numérica tenga que ser lineal. Dos variables numéricas pueden tener una relación no lineal. Por ejemplo, la edad en años y las horas pasadas en un puesto de trabajo pueden ser medidas ambas a nivel numérico, pero dado que tanto los niños como los jubilados pasan poco o ningún tiempo en el trabajo, la correlación lineal entre estas dos variables será probablemente muy baja.

El escalamiento óptimo puede detectar relaciones no lineales y producir correlaciones máximas entre variables. Los cuatro procedimientos de escalamiento óptimo antes definidos amplían el ámbito de aplicación de las técnicas estadísticas clásicas de Análisis de Componentes Principales (ACP) y de Análisis de Correlación Canónica (ACC), para acomodar variables de niveles mixtos de medida. Si todas las variables del análisis fuesen numéricas y las relaciones entre las variables lineales, entonces deberían emplearse los procedimientos estadísticos estándares basados en la correlación y no habría necesidad de utilizar los procedimientos de escalamiento óptimo. Sin embargo, si las variables de análisis tienen niveles mixtos de medida, o si se sospecha que existen relaciones no lineales entre algunos pares de variables, entonces debería utilizarse el procedimiento de escalamiento óptimo.

En el escalamiento óptimo, el usuario especifica el tipo de medida de cada variable, diferenciando el nivel de medida de cada una de las variables del análisis, permitiendo así la búsqueda de soluciones con el fin de que las variables elegidas por el modelo se ajusten bien a los datos. El escalamiento óptimo también revelará relaciones no lineales. Esto se hace de modo exploratorio, en contraposición con las pruebas de hipótesis estándar en el contexto de las suposiciones distributivas, tales como la normalidad y la linealidad de la regresión de las variables originales.

El escalamiento óptimo proporciona un conjunto de puntuaciones óptimas (o cuantificaciones de categorías), para las categorías de cada variable. Las puntuaciones óptimas se asignan a las categorías de cada variable, basadas en el criterio de optimización del procedimiento en uso. A diferencia de los valores originales de las variables nominales u ordinales del análisis, estas puntuaciones tienen propiedades métricas, por lo que éstas técnicas se describen frecuentemente como una forma de cuantificación de datos cualitativos, que también incluyen técnicas como el escalamiento no métrico multidimensional (disponible en el procedimiento ALSCAL). Las cuantificaciones de las categorías de cada variable pueden representarse sobre un plano bidimensional o, incluso, en un plano tridimensional, siendo su yuxtaposición en el mismo gráfico útil para revelar patrones de asociación entre variables.

CORRELACIÓN CANÓNICA NO LINEAL

El objetivo del análisis no lineal de la correlación canónica es analizar las relaciones entre dos o más grupos de variables. En el análisis de correlación canónica hay dos grupos de variables numéricas: por ejemplo, un grupo de variables, formado por los ítems demográficos en un grupo de encuestados, y un grupo de variables, con respuestas a un grupo de ítems de actitud. El análisis de correlación canónica estándar (que se estudiará en un tema posterior) es una técnica estadística que busca una combinación lineal de un grupo de variables y una combinación lineal de un segundo grupo de variables correlacionadas al máximo.

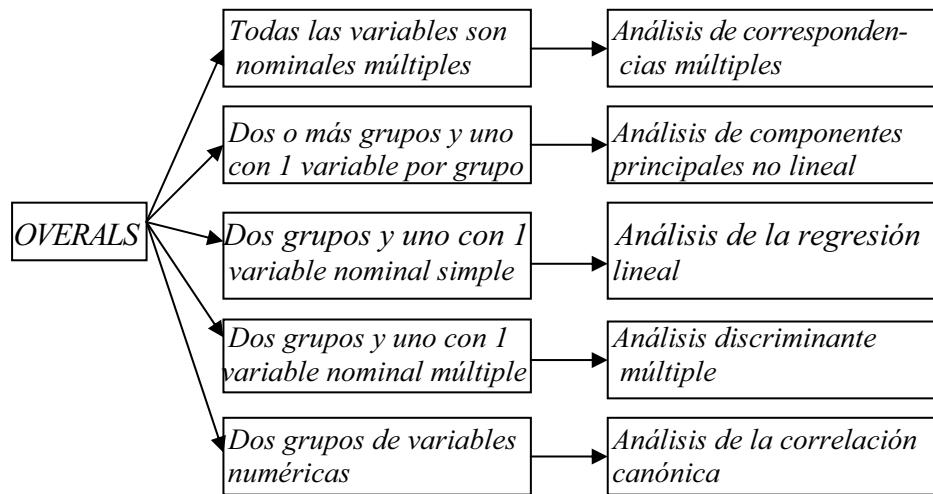
Dado este hecho, el análisis de correlación canónica puede buscar los grupos independientes subsiguientes de combinaciones lineales, hasta un máximo igual al número de variables del grupo más pequeño.

El procedimiento OVERALS generaliza el Análisis de Correlación Canónica al permitir lo siguiente:

- Existencia de 2 o más grupos de variables, por lo que no se está restringido a 2 grupos de variables, como ocurre en las ejecuciones más populares del análisis de la correlación canónica.
- Las variables de OVERALS pueden ser tanto variables nominales, como ordinales o numéricas.
- El procedimiento OVERALS determina la similitud entre grupos, comparando simultáneamente las combinaciones lineales de las variables en cada grupo con las puntuaciones de los objetos.

El análisis de correlación canónica no lineal o análisis de correlación canónica categórico mediante escalamiento óptimo tiene como propósito final determinar la similitud entre los conjuntos de variables categóricas. El análisis de correlación canónica estándar es una extensión de la regresión múltiple, en la que el segundo conjunto no contiene una única variable de respuesta, sino varias. El objetivo es explicar el máximo posible de la varianza sobre las relaciones existentes entre dos conjuntos de variables numéricas en un espacio de pocas dimensiones. Inicialmente, las variables de cada conjunto se combinan linealmente de forma que las combinaciones lineales tengan una correlación máxima entre sí. Una vez dadas estas combinaciones, se establece que las combinaciones lineales subsiguientes no estén correlacionadas con las combinaciones anteriores y que también tengan la mayor correlación posible.

El análisis no lineal de correlación canónica está relacionado con otros procedimientos. Si para dos grupos de variables se tiene que uno de los grupos presenta una variable nominal declarada como nominal simple, los resultados de OVERALS se pueden interpretar de un modo similar al Análisis de Regresión Múltiple. Si para dos grupos de variables se considera a la variable nominal múltiple, OVERALS es una alternativa al Análisis Discriminante. Si para más de 2 grupos de variables se tiene que todas las variables son numéricas, OVERALS equivale al Análisis de Correlación Canónica estándar. Si para más de 2 grupos de variables cada grupo tiene una sola variable, OVERALS equivale a componentes principales categórico (CATPCA). Si para más de 2 grupos de variables todas las variables son nominales múltiples, OVERALS equivale al análisis de correspondencias múltiples HOMALS. El cuadro siguiente resume las relaciones anteriores:



La aproximación por escalamiento óptimo expande el análisis estándar de tres formas decisivas. Primera: OVERALS permite más de dos conjuntos de variables. Segunda: las variables se pueden escalar como nominales, ordinales o numéricas. Como resultado, se pueden analizar relaciones no lineales entre las variables. Finalmente, en lugar de maximizar las correlaciones entre los conjuntos de variables, los conjuntos se comparan con un conjunto de compromiso desconocido definido por las puntuaciones de los objetos.

ANÁLISIS DE COMPONENTES PRINCIPALES CATEGÓRICAS

Ya hemos visto en un capítulo anterior que el análisis de componentes principales estándar es una técnica estadística que transforma linealmente un paquete original de variables en otro sustancialmente más pequeño de variables incorreladas, que representa la mayoría de la información del grupo original de variables. El objetivo del análisis de componentes principales es reducir la dimensionalidad del conjunto de datos originales mientras se contabiliza tanto como sea posible la variación en el grupo original de variables. En el análisis de componentes principales se pueden asignar puntuaciones de los componentes a los objetos del análisis. Los gráficos de las puntuaciones de los componentes revelan patrones entre los objetos del análisis y pueden revelar objetos inusuales en los datos.

El análisis estándar de componentes principales asume que todas las variables del análisis se miden a escala numérica, y que las relaciones entre los pares de las variables son lineales. El análisis de componentes principales categóricas extiende esta metodología para permitir la ejecución del análisis de componentes principales en cualquier mezcla de variables nominales, ordinales y numéricas.

El objetivo del análisis de componentes principales categóricas sigue siendo contabilizar la mayor variación posible en los datos, dada la dimensionalidad especificada del análisis. Para las variables nominales y ordinales del análisis, el programa calcula las puntuaciones óptimas para las categorías.

Si todas las variables se declaran numéricas, el análisis no lineal de componentes principales (CATPCA) equivale al análisis de componentes principales estándar utilizando el análisis factorial. Si todas las variables se declaran nominales múltiples, el análisis no lineal de componentes principales equivale al análisis de correspondencias múltiples (HOMALS) ejecutado sobre las mismas variables. Por tanto, CATPCA es un tipo de HOMALS con algunas variables ordinales o numéricas.

El análisis de componentes principales categórico se conoce también por el acrónimo CATPCA, del inglés Categorical Principal Components Analysis. Este procedimiento cuantifica simultáneamente las variables categóricas a la vez que reduce la dimensionalidad de los datos. El objetivo de los análisis de componentes principales es la reducción de un conjunto original de variables en un conjunto más pequeño de componentes no correlacionados que representen la mayor parte de la información encontrada en las variables originales. La técnica es más útil cuando un extenso número de variables impide una interpretación eficaz de las relaciones entre los objetos (sujetos y unidades). Al reducir la dimensionalidad, se interpreta un pequeño número de componentes en lugar de un extenso número de variables.

El análisis típico de componentes principales asume relaciones lineales entre las variables numéricas. Por otra parte, la aproximación por escalamiento óptimo permite escalar las variables a diferentes niveles. Las variables categóricas se cuantifican de forma óptima en la dimensionalidad especificada. Como resultado, se pueden modelar relaciones no lineales entre las variables.

CONCEPTO DE ESCALAMIENTO MULTIDIMENSIONAL

El escalamiento multidimensional trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos. Esto se logra asignando las observaciones a posiciones específicas en un espacio conceptual (normalmente de dos o tres dimensiones) de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades dadas. En muchos casos, las dimensiones de este espacio conceptual son interpretables y se pueden utilizar para comprender mejor los datos. Si las variables se han medido objetivamente, puede utilizar el escalamiento multidimensional como técnica de reducción de datos (el procedimiento escalamiento multidimensional permitirá calcular las distancias a partir de los datos multivariados, si es necesario). El escalamiento multidimensional puede también aplicarse a valoraciones subjetivas de disimilaridad entre objetos o conceptos.

Además, el procedimiento escalamiento multidimensional puede tratar datos de disimilitud procedentes de múltiples fuentes, como podrían ser múltiples evaluadores o múltiples sujetos evaluados por un cuestionario.

Según Luque (2000), el escalamiento multidimensional (EMD) se origina en psicología como una respuesta a la necesidad de relacionar la intensidad física de ciertos estímulos con su intensidad subjetiva. Torgerson (1958) es considerado como uno de sus principales precursores, contribuyendo decisivamente a la clasificación y utilización de estos métodos. Este autor fue el primero en proponer una generalización del escalamiento. Pronto surgieron nuevos modelos y métodos que paulatina y sistemáticamente fueron cubriendo un amplio abanico de demandas realizadas desde diferentes campos de investigación como la Psicología, la Educación, Sociología, las Ciencias Políticas, la Economía y, por supuesto, el Marketing. Un factor que favoreció su desarrollo fue la evolución experimentada por los equipos informáticos y el software a partir de los años cincuenta. Ello permitió el desarrollo de numerosos algoritmos de escalamiento multidimensional (EMD) materializados en programas de amplia difusión a nivel mundial (KYST, INDSCAL, SINDSCAL, MULTISCALE, ALSCAL, PREFMAP, etc.). Incluso, paquetes estadísticos tan populares como SPSS, STATISTICA y SYSTAT tienen implementados sus propios programas de EMD.

Comúnmente, el escalamiento multidimensional se clasifica dentro de los métodos de interdependencia y es un procedimiento que permite al investigador determinar la imagen relativa percibida de un conjunto de objetos (empresas, productos, ideas u otros objetos sobre los que los individuos desarrollan percepciones). Es decir, el aspecto característico de este procedimiento es que proporciona una representación gráfica en un espacio geométrico de pocas dimensiones que permite comprender cómo los individuos perciben objetos y qué esquemas, generalmente ocultos, están detrás de esa percepción. En estos espacios, los objetos adoptan la forma de puntos y la proximidad entre ellos refleja la analogía existente entre los mismos. La interpretación de las dimensiones depende del conocimiento que se tenga acerca de esos estímulos y se realiza de forma similar a como se haría con un análisis factorial clásico o un análisis de correspondencias.

El objetivo del escalamiento multidimensional es transformar los juicios de similitud o preferencias llevados a cabo por una serie de individuos en distancias susceptibles de ser representadas en un espacio multidimensional. Así, por ejemplo, si un conjunto de individuos opina que los objetos A y B son los dos más parecidos de entre un conjunto de objetos, el escalamiento multidimensional posicionaría A y B de modo que la distancia entre ambos sea la menor de las existentes entre cada par de objetos. El mapa perceptual resultante muestra la posición relativa del conjunto de objetos sobre los que se centra el estudio. El tipo de datos que hay que recabar son juicios de similitud, disimilitud o preferencia que los sujetos encuestados manifiestan en relación con todas las posibles combinaciones de pares de objetos a investigar. La aplicación de esta técnica no requiere un conocimiento previo de los atributos que los sujetos utilizan al emitir sus juicios.

Tampoco se precisa un nivel de medida muy restrictivo para operativizar los juicios que se realicen. El escalamiento multidimensional está basado en la comparación de objetos, admitiendo, que cualquier objeto está formado tanto por dimensiones objetivas como por dimensiones subjetivas o perceptuales.

Las dos principales repercusiones para la investigación de esta importante diferenciación entre atributos objetivos y percibidos son que las dimensiones percibidas por los consumidores no tienen por qué coincidir con las dimensiones objetivas asumidas como relevantes por el investigador y las evaluaciones de dichas dimensiones (aun en el caso de que las dimensiones percibidas coincidan con las objetivas) pueden no ser independientes o no coincidir con los valores objetivos.

Es necesario hacer hincapié sobre la precaución necesaria en la interpretación de los resultados de este tipo de análisis. Dicha interpretación constituye más un arte que una ciencia, es decir, no existen reglas fijas para llevarla a cabo. Es por ello que el analista debiera resistirse a la tentación de permitir que sus propias percepciones afecten a la interpretación de las dimensiones percibidas por los individuos encuestados. En definitiva, el EMD es una herramienta muy útil cuando se pretende investigar objetos para los que el conocimiento está poco organizado y los esquemas perceptuales son poco o nada conocidos.

Mapas perceptuales

Las técnicas de elaboración de mapas perceptuales, y en particular el escalamiento multidimensional, resultan especialmente apropiadas para la satisfacción de un primer objetivo consistente en la identificación de dimensiones no reconocidas susceptibles de afectar al comportamiento. Un segundo objetivo sería la obtención de evaluaciones comparativas de objetos en aquellos casos en los que las bases de comparación son desconocidas o no están definidas. En el EMD no es necesario que el investigador ni los individuos entrevistados especifiquen los atributos de comparación. No obstante, el analista sí tiene que especificar los objetos a comparar y asegurarse de que éstos comparten una base común de comparación.

Cuando se define el análisis es necesario, en primer lugar, asegurarse de que todos los objetos relevantes (empresas, productos, servicios u otros), y sólo éstos, son incluidos. Además hay que cerciorarse de que éstos sean comparables entre sí, ya que el escalamiento multidimensional es una técnica de posicionamiento relativo. La relevancia de un objeto viene determinada por los objetivos perseguidos por el investigador. En segundo lugar, hay que decidir el número de objetos a evaluar. Así, se ha de buscar un equilibrio entre un número reducido de objetos que facilite la evaluación por el entrevistado y un número mayor que permita la obtención de una solución estable. A modo de orientación, el número de objetos debe superar en cuatro veces el de las dimensiones.

El número de objetos afecta también a la obtención de un nivel aceptable de ajuste. En muchas ocasiones, la utilización de un número de objetos inferior al sugerido para determinada dimensionalidad provoca una supravaloración de la bondad del ajuste.

En cuanto a la *Elección del tipo de datos*, el investigador debe optar entre la obtención de datos de similitud o de preferencias. Los mapas perceptuales basados en similitudes representan el parecido entre los atributos de los objetos, así como las dimensiones perceptuales empleadas en la comparación, si bien no reflejan las preferencias de los individuos respecto a los objetos ni sus determinantes. Los mapas perceptuales basados en datos de preferencias sí que reflejan qué objetos son preferidos, si bien las posiciones resultantes no tienen por qué coincidir con las basadas en juicios de similitud, ya que los individuos encuestados pueden en cada caso basar sus valoraciones en dimensiones completamente distintas.

En cuanto a la *Elección del tipo de análisis*, el investigador puede generar el *output* sujeto por sujeto, generando tantos mapas como sujetos han sido entrevistados, lo que se conoce como análisis desagregado. Sin embargo, las técnicas de EMD permiten también combinar las respuestas de los individuos entrevistados para generar un menor número de mapas perceptuales, mediante un proceso de análisis agregado, previo o posterior al escalamiento multidimensional de los datos ofrecidos por los sujetos. El modo de agregación más simple consiste en encontrar una *Devaluación media* para cada grupo de individuos (formado, por ejemplo, mediante un análisis *cluster*) y obtener una solución agregada única a partir de ésta. También podemos utilizar el modelo INDSCAL (*Individual Differences Scaling*) y sus variantes que permiten realizar un análisis desagregado especializado. La elección entre análisis agregado o desagregado depende una vez más de los objetivos del estudio. Si el objetivo es conocer las evaluaciones globales de los objetos y las dimensiones empleadas en sus evaluaciones, el análisis agregado resulta más adecuado, mientras que si el objetivo es conocer las variaciones entre los individuos, el enfoque desagregado es el más adecuado.

En cuanto a la *Elección del método de análisis* las opciones que se presentan son: métodos no métricos y métricos. Los métodos no métricos, llamados así por el carácter no métrico de los datos de entrada (comúnmente generados mediante la ordenación de pares de objetos), resultan más flexibles al no asumir ningún tipo específico de relación entre la distancia calculada y la medida de similitud. Sin embargo, es más probable que resulten en soluciones degeneradas o no óptimas. Los métodos métricos se distinguen por el carácter métrico tanto de los datos de entrada como de los resultados. Este supuesto nos permite reforzar la relación entre la dimensionalidad de la solución final y los datos iniciales. Cabe suponer que la solución mantiene el carácter métrico de los datos iniciales.

El escalamiento multidimensional requiere que el investigador acepte algunos supuestos relacionados con la percepción, como por ejemplo que cada individuo entrevistado percibirá cada estímulo según unas dimensiones (aunque la mayoría de las personas juzgan en términos de un número limitado de dimensiones o características). Como caso práctico podemos fijarnos en que algunas personas evalúan un coche en términos de potencia y aspecto, en tanto que otras no consideran estos factores y lo juzgan en términos de coste y confort interior. Otro supuesto que ha de aceptar el investigador es que no necesariamente todos los individuos otorgan la misma importancia a determinada dimensión o atributo. También es un supuesto a aceptar por el investigador que los juicios acerca de un estímulo no tienen por qué mantenerse estables en el tiempo ni en lo relativo a sus dimensiones ni en cuanto a la importancia otorgada a éstas.

Es lógico esperar que el escalamiento multidimensional represente espacialmente las percepciones de modo que sea posible examinar cualquier relación subyacente común. El propósito de estas técnicas no es únicamente el de conocer individualmente a las personas entrevistadas, sino también identificar las percepciones y las dimensiones de evaluación compartidas por los individuos que componen la muestra.

Solución, ajuste y preferencias en el escalamiento multidimensional

Se trata de determinar la posición de cada objeto en el espacio perceptual de modo que los juicios de similitud expresados por los individuos entrevistados se reflejen lo más fielmente posible. Los programas de escalamiento multidimensional siguen un ***procedimiento común para la determinación de las posiciones óptimas***, que se resume en varios pasos

El primer paso es la *Selección de una configuración inicial de los estímulos* según la dimensionalidad inicial deseada. Existen distintas opciones para obtener una configuración inicial. Las dos más empleadas consisten en utilizar una configuración desarrollada por el propio investigador sobre la base de trabajos de investigación previos o bien una configuración generada seleccionando puntos pseudoaleatorios a partir de una distribución normal multivariante.

Un segundo paso sería el *Cálculo de las distancias entre los puntos* representativos de los estímulos y comparación de las relaciones (observadas versus derivadas) mediante una medida de ajuste o *Stress*.

Si el indicador de ajuste no alcanza un valor mínimo previamente fijado por el investigador, un tercer paso sería *Encontrar una nueva configuración* para la que el indicador de ajuste sea mejor. El programa/algoritmo determinará las direcciones que producen las mayores mejoras en el ajuste y moverá poco a poco los puntos en dichas direcciones.

En un cuarto paso, el programa realizará una *Evaluación de la nueva configuración y la ajustará* hasta que se logre obtener un nivel satisfactorio de ajuste.

Un quinto y último paso sería la *Reducción de la dimensionalidad de la configuración actual* y repetición del proceso hasta lograr obtener aquella configuración que, con la menor dimensionalidad posible, presente un nivel de ajuste aceptable.

El analista debe preocuparse de obtener varias soluciones con diferente número de dimensiones y elegir entre ellas sobre la base de tres criterios fundamentales: su nivel de ajuste a los datos, su interpretabilidad y su replicabilidad.

En relación con el **nivel de ajuste**, se trata de calcular una medida de *stress*, que indica la proporción de varianza de los datos originales no recogida por el modelo de escalamiento multidimensional. Esta medida varía según el tipo de programa y el tipo de datos que se estén analizando. En cualquier caso, el *stress* mejora a medida que se consideran más dimensiones. Entre las **medidas de la bondad del ajuste más usuales** se tienen la medida *Stress* de Kruskal basada en las disparidades (datos óptimamente escalados) y las distancias, la medida *S-stress* utilizada por el algoritmo ALSCAL, la medida *RSQ* y el *Coeficiente de alienación*.

Para la medida *Stress* hay dos expresiones:

$$S_1 = \sqrt{\frac{\sum_{i} \sum_{j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i} \sum_{j} d_{ij}^2}} \quad S_2 = \sqrt{\frac{\sum_{i} \sum_{j} (d_{ij} - \hat{d}_{..})^2}{\sum_{i} \sum_{j} (\hat{d}_{ij} - \hat{d}_{..})^2}} \quad \hat{d}_{..} = \frac{1}{n^2} \sum_i \sum_j \hat{d}_{ij}$$

donde $\hat{d}_{..}$ es la media aritmética de las distancias estimadas y d_{ij} son las distancias originales entre objetos.

Suele utilizarse la primera expresión cuando la matriz de disimilitudes es simétrica.

Para la medida *S-stress* hay también dos expresiones:

$$SS_1 = \sqrt{\frac{\sum_{i} \sum_{j} (d_{ij}^2 - \hat{d}_{ij}^2)^2}{\sum_{i} \sum_{j} (d_{ij}^2)^2}} \quad SS_2 = \sqrt{\frac{\sum_{i} \sum_{j} (d_{ij}^2 - \hat{d}_{..}^2)^2}{\sum_{i} \sum_{j} (\hat{d}_{ij}^2 - \hat{d}_{..}^2)^2}} \quad \hat{d}_{..} = \frac{1}{n^2} \sum_i \sum_j \hat{d}_{ij}$$

Vemos que también en este caso existen dos fórmulas como función a minimizar para estimar las coordenadas de los estímulos. Suele utilizarse la primera expresión cuando los datos originales son disimilitudes, y la segunda cuando son preferencias.

La medida *RSQ* es el índice de correlación al cuadrado R^2 o RSQ que mide la proporción de varianza de las disparidades correspondiente al escalamiento multidimensional, es decir, mide lo bien que se ajustan los datos originales al modelo de escalamiento multidimensional. Suele ser aceptable a partir de 0,6.

El coeficiente de alienación κ de Guttman es también una medida de bondad de ajuste en análisis no métricos. Suele definirse en función del coeficiente de monotonicidad μ como sigue:

$$\mu = \frac{\sum_{i} \sum_{j} d_{ij} \hat{d}_{ij}}{\sqrt{\left(\sum_{i} \sum_{j} d_{ij}^2 \right) \left(\sum_{i} \sum_{j} \hat{d}_{ij}^2 \right)}} \quad \kappa = \sqrt{1 - \mu^2}$$

La medida κ cuantifica la maldad del ajuste, con lo que interesará que su valor sea lo menor posible.

En general, suelen rechazarse soluciones con un *stress* superior a 0,10, a menos que se trate de soluciones unidimensionales. Sin embargo, si los datos contienen altos niveles de error muestras o de medida, cabe la posibilidad de aceptar valores de *stress* superiores a 0,10. En general, se suele considerar que 0,05 es un nivel aceptable de *stress* y que valores por debajo de 0,01 indican un nivel muy bueno de ajuste. La *replicabilidad* es un criterio aplicable únicamente a aquellas situaciones en las que se cuenta con dos o más submuestras. El objetivo es retener aquellas dimensiones que aparezcan de forma consistente para las distintas submuestras. Si se aplica el escalamiento multidimensional por separado a las distintas submuestras y existen t dimensiones que aparecen en las distintas soluciones, la solución final debe contener exactamente estas t dimensiones. Lógicamente, las distintas submuestras deben provenir de la misma población. El criterio de la *interpretabilidad* requiere de cierto juicio subjetivo por parte del analista. En este sentido, una solución con una dimensionalidad superior será preferible a otra con una dimensionalidad menor si existen ciertos atributos importantes de los estímulos que aparecen en la primera y que no son recogidos por la segunda. En caso contrario, dada su sencillez, una solución con una dimensionalidad menor será preferible.

En cuanto a la *incorporación de preferencias en el escalamiento multidimensional*, hay que tener en cuenta que los mapas perceptuales pueden también derivarse de datos de preferencias. El objetivo es, dada una configuración para un conjunto de objetos, determinar la combinación de características preferida. Así se desarrolla un espacio conjunto donde se representan tanto los objetos (estímulos) como los sujetos (puntos ideales). Para ello es preciso asumir el supuesto de homogeneidad en las percepciones de los individuos en relación con el conjunto de objetos. Esto permite que todas las diferencias sean atribuidas a las preferencias y no a las diferencias perceptuales. La incorporación de preferencias da lugar a un resultado de enorme interés, el *punto ideal* para cada individuo entrevistado. Identificar la posición de un objeto ideal en el mapa perceptual implica localizar la combinación preferida de atributos percibidos. Asumimos que la posición de este punto ideal (en relación con el resto de objetos representados en el mapa perceptual) define la preferencia relativa, de modo que aquellos objetos más alejados de dicho punto serán los menos preferidos.

Los dos procedimientos empleados generalmente para la determinación de los puntos ideales son la *estimación explícita* y la *estimación implícita*. La estimación explícita toma como base las respuestas directas de los individuos, solicitándoles que evalúen un producto ideal hipotético en relación con los mismos atributos empleados para evaluar el resto de objetos. Esto supone una serie de problemas, ya que los individuos tienden a situar su objeto ideal en los extremos de las valoraciones explícitas empleadas o a considerarlo similar al objeto preferido. Además, el individuo debe razonar, no en términos de similitudes, sino de preferencias, lo que a menudo resulta difícil cuando se trata de objetos relativamente desconocidos. Esas dificultades suelen llevar a los investigadores a realizar estimaciones implícitas de los puntos ideales, a través distintos procedimientos. El supuesto básico que subyace a la mayoría de dichos procedimientos es que las medidas derivadas de las posiciones espaciales de los puntos ideales son consistentes con las preferencias de los individuos. Srinivasan y Shocker (1973) asumen que el punto ideal para un conjunto de pares de estímulos es aquél que en un menor número de casos deja de cumplir la restricción de encontrarse más cerca del más preferido dentro de cada par.

El posicionamiento implícito de los puntos ideales a partir de los datos de preferencia puede llevarse a cabo en primer lugar mediante *análisis internos de los datos de preferencia*, que implica el desarrollo de mapas espaciales en los que simultáneamente se representan estímulos y sujetos (mediante puntos o vectores) partiendo únicamente de los datos de preferencia. Estos métodos suponen que las posiciones de los objetos se calculan mediante un desdoblamiento de los datos de preferencias correspondientes a cada individuo y también suponen que los resultados reflejan dimensiones perceptuales que son ponderadas para predecir las preferencias. Generalmente emplean una representación sectorial del punto ideal, mientras que los métodos externos pueden estar basados tanto en representaciones sectoriales como puntuales.

El posicionamiento implícito de los puntos ideales a partir de los datos de preferencia puede llevarse a cabo en segundo lugar mediante análisis externos de los datos de preferencia. El análisis externo de datos de preferencia consiste en ajustar los puntos ideales (basados en datos de preferencia) a un espacio desarrollado a partir de datos de similitudes obtenidos de los mismos sujetos. Por ejemplo, podrían desarrollarse mapas individuales a partir de datos de similitudes, examinar dichos mapas en busca de rasgos comunes y representar los datos de preferencias en relación con los grupos de individuos identificados. Por tanto, para poder realizar un análisis externo el investigador debe contar con datos de preferencias y datos de similitudes.

El análisis externo suele ser preferible en la mayoría de situaciones dadas las dificultades de cálculo de los procedimientos de análisis interno y la confusión entre diferencias en las percepciones. Además, la importancia de las dimensiones percibidas puede cambiar cuando pasamos de un espacio perceptual a un espacio de preferencias.

En la representación de mapas perceptuales basados en datos de preferencia mediante puntos ideales (*representación puntual*), el orden de preferencias puede extraerse a partir de las distancias euclidianas que separan al punto ideal del resto de puntos representativos de los distintos objetos. En este caso, estaríamos asumiendo que la dirección de la distancia carece de importancia y únicamente consideraríamos la distancia relativa. Pero las preferencias pueden también representarse por medio de un vector (*representación vectorial*). Para calcular las preferencias bajo este enfoque, se trazan líneas perpendiculares (proyecciones) desde el objeto hacia el vector. Las preferencias son mayores en el sentido indicado por el vector. Éstas pueden derivarse directamente a partir del orden de las proyecciones.

Interpretación y validación de los resultados

Existen procedimientos subjetivos para la interpretación de los resultados del escalamiento multidimensional. Este tipo de interpretación supone siempre un juicio por parte del investigador o del entrevistado, y en muchos casos esto constituye una solución para el problema que nos ocupa. Un modo bastante simple aunque efectivo de etiquetar las dimensiones del mapa perceptual consiste en pedir a los individuos entrevistados que, tras una inspección visual del mapa resultante, interpreten subjetivamente la dimensionalidad del mismo. También cabe pedir a un conjunto de expertos que evalúen e identifiquen las dimensiones. Si bien no se persigue relacionar cuantitativamente las dimensiones con los atributos, este enfoque será el más adecuado en aquellos casos con dimensiones de carácter intangible o de contenido afectivo o emocional.

De forma similar, el propio investigador puede describir las dimensiones en términos de dimensiones conocidas (objetivas). De este modo se establece directamente una correspondencia entre las dimensiones objetivas y perceptores. Estaríamos ante los procedimientos objetivos para la interpretación de los resultados del escalamiento multidimensional. Como complemento de los procedimientos subjetivos, el investigador cuenta con una serie de métodos más formalizados. El método más empleado, PROFIT (*Property Fitting*), recoge las puntuaciones respecto a los atributos de cada objeto y encuentra la mejor correspondencia entre cada atributo y el espacio perceptual derivado. El objetivo es identificar los atributos determinantes de los juicios de similitud realizados por los individuos entrevistados. Este método ofrece una medida de ajuste para cada atributo, así como su correspondencia con las dimensiones. El analista puede entonces determinar qué atributos describen mejor las posiciones perceptuales y son más ilustrativas de las dimensiones. La necesidad de una correspondencia entre los atributos y las dimensiones definidas es menor en el caso de obtener resultados métricos, ya que las dimensiones pueden rotarse libremente sin que ello afecte a las posiciones relativas de los objetos.

Tanto si se adopta un procedimiento subjetivo como si se opta por uno objetivo, el investigador debe recordar que es habitual que una dimensión represente a más de un atributo. La mejor alternativa para apoyar la interpretación de las dimensiones consiste en utilizar los datos referidos a atributos. Sin embargo, existe el riesgo de que el analista no considere todos los atributos relevantes. En todo caso, la interpretación debe hacerse atendiendo a la existencia de agrupaciones u ordenaciones significativas de los estímulos. Una agrupación significativa de los estímulos es un conjunto de estímulos que aparecen juntos en una determinada región del espacio multidimensional resultante y que poseen ciertas características comunes.

Por otra parte, una ordenación significativa de los estímulos es aquélla que dispone a los estímulos según su mayor o menor contenido de alguno de los atributos relevantes. Por tanto, la interpretación de la configuración final consistirá en determinar las características comunes de los estímulos que forman agrupaciones significativas y en determinar las características que dan lugar a ordenaciones significativas.

En cuanto a la **validación de los resultados**, dada la naturaleza inferencial del escalamiento multidimensional, la validación de los resultados debe dirigirse a asegurar su generalización a otros objetos y a otros individuos de la población, lo que no siempre es fácil. La única alternativa de comparación de resultados es la posición relativa de los objetos. Las dimensiones subyacentes no tienen ninguna base de comparación. Si las posiciones varían, el investigador no puede determinar si los objetos son percibidos de un modo distinto o si las dimensiones perceptuales han variado (o ambas cosas a la vez). No se han desarrollado métodos sistemáticos de comparación que hayan sido integrados en los programas estadísticos. ¿Cuáles son las opciones disponibles? El enfoque más directo consiste en realizar una división de la muestra u obtener distintas muestras y comparar los resultados de éstas.

En ambos casos, el investigador debe encontrar un medio para comparar los resultados. Frecuentemente dicha comparación se realiza visualmente o mediante una simple correlación de las coordenadas. Otro modo de validar los resultados consiste en aplicar métodos de composición (*cluster, factorial* o *correspondencias*) y de descomposición (escalamiento multidimensional) a la misma muestra de individuos. Con los métodos de descomposición se interpretan las dimensiones resultantes para identificar los atributos clave, y después se aplica uno o más métodos de composición (en especial el análisis de correspondencias) para confirmar los resultados.

MODELOS DE ESCALAMIENTO MULTIDIMENSIONAL

Existen varios modelos prácticos de escalamiento multidimensional, cada uno de ellos utilizable en un determinado tipo de situación. De todas formas, la elección del método de escalamiento dependerá del tipo de datos de que se dispone y del tipo de información que se desee extraer de los datos.

De modo general, el modelo de escalamiento multidimensional toma como entrada habitual una matriz cuadrada de proximidades Δ de tamaño (n,n) donde n es el número de estímulos. Para $n = 4$ la matriz Δ es la siguiente:

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix}$$

A partir de esta matriz de proximidades, el modelo de escalamiento multidimensional proporciona como solución una matriz rectangular X de tamaño (n,m) donde n es el número de estímulos y m es el número de dimensiones. Cada valor x_{ia} de la matriz X corresponde a la coordenada del estímulo i en la dimensión a . Los estímulos suelen representarse en un espacio de dos o tres dimensiones como mucho en la mayoría de los casos. Es rara la dimensionalidad superior a 4 y la mayoría de los programas tienen el límite superior en 6 dimensiones. Para 2 dimensiones y 4 estímulos, la matriz X sería la siguiente:

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix}$$

Cada fila $[x_{i1}, x_{i2}]$ de la matriz X contiene las coordenadas del estímulo i en los ejes de coordenadas X e Y que delimitan el espacio bidimensional. A partir de la matriz X es posible situar los n estímulos en el espacio asignándoles los valores de coordenadas correspondientes. También es posible utilizar la matriz X para calcular las distancias entre dos estímulos i y j cualquiera aplicando la fórmula general de la distancia de Minkowski:

$$d_{ij} = \left[\sum_{a=1}^m (x_{ia} - x_{ja})^p \right]^{\frac{1}{p}} \quad 1 \leq p \leq \infty$$

Generalmente p es 2 y corresponde a la métrica euclídea.

La estimación de las distancias correspondientes a todos los estímulos proporciona una nueva matriz D, que en nuestro caso sería la siguiente:

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix}$$

La solución proporcionada por el escalamiento multidimensional debe de ser tal que exista la máxima correspondencia entre las proximidades entre estímulos proporcionadas en la matriz Δ y las distancias entre estímulos obtenidas en la matriz D.

Modelo de escalamiento métrico

Los modelos de escalamiento parten de una función de representación de las proximidades estimadas por los sujetos en forma de distancias entre objetos: $d_{ij} \rightarrow f(\delta_{ij})$. En el caso del modelo métrico (también llamado *clásico*), la relación planteada generalmente entre proximidades y distancias es de tipo lineal: $d_{ij} = a + b \delta_{ij}$, aunque muchas variantes del modelo métrico admiten también transformaciones potenciales, logarítmicas o polinómicas de cualquier grado.

El modelo métrico parte de una matriz cuadrada $D(n \times n)$ de distancias entre n objetos, a partir de la cual es posible derivar una matriz $B(n \times n)$ de productos escalares entre vectores. A su vez, es posible descomponer esta matriz B en el producto XX' , donde $X(n \times r)$ es una matriz rectangular de coordenadas de los n objetos en un espacio de r dimensiones. Para llevar a cabo el procedimiento, el primer paso consistirá en la conversión de la matriz de proximidades $\Delta(n \times n)$ en una matriz de distancias.

El modelo de distancia euclídea exige el cumplimiento de tres axiomas, que son los siguientes:

- *Axioma de no-negatividad.* Las distancias son valores no negativos ($d_{ij} \geq d_{ii} = 0$).
- *Axioma de simetría.* La distancia entre dos objetos i y j es simétrica ($d_{ij} = d_{ji}$).
- *Axioma de desigualdad triangular.* La distancia entre dos objetos i y j no puede ser mayor que la suma de las distancias de i y j a un tercer objeto k ($d_{ij} \leq d_{ik} + d_{kj}$).

Es necesario, por tanto, que Δ satisfaga los tres axiomas para obtener una matriz de distancias. Resulta relativamente sencillo satisfacer los dos primeros, utilizando únicamente valores positivos o cero para las proximidades, y empleando únicamente una de las mitades de la matriz de proximidades. Sin embargo, para cumplir el tercer axioma es necesario calcular un valor que, sumado a las proximidades originales, nos permitirá cumplir el axioma. Torgerson determinó el valor mínimo de esta *constante aditiva* (c) para toda terna de objetos i, j, k como el valor máximo de las diferencias entre ternas de proximidades:

$$c_{\min} = \max_{(i,j,k)} (\delta_{ij} - \delta_{ik} - \delta_{kj})$$

Una vez sumado el valor de c a todos los elementos de la matriz de proximidades ya tenemos la matriz D de distancias entre objetos. A continuación, transformamos la matriz D en una matriz B de productos escalares entre vectores. Los elementos b_{ij} de esta nueva matriz se crean a partir de los elementos d_{ij} de D mediante la siguiente transformación:

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{ii}^2 - d_{jj}^2 + d^2)$$

Donde:

$$d_{ii}^2 = \frac{1}{n} \sum_j^n d_{ij}^2 \quad (\text{distancia cuadrática media por fila}).$$

$$d_{jj}^2 = \frac{1}{n} \sum_i^n d_{ij}^2 \quad (\text{distancia cuadrática media por columna}).$$

$$d^2 = \frac{1}{n^2} \sum_i^n \sum_j^n d_{ij}^2 \quad (\text{distancia cuadrática media de la matriz}).$$

Para finalizar, sólo es necesario hallar la matriz de coordenadas X , tal que $B = XX'$. Aunque la mayoría de programas de MDS actuales se valen de complejos algoritmos para determinar los valores de X , esta descomposición puede llevarse a cabo mediante componentes principales (Hotelling, 1933).

El método de componentes principales nos permite descomponer una matriz cuadrada B en un producto de dos matrices, $Q(n \times n)$ y $\Lambda(n \times n)$, siendo Q ortonormal ($Q' Q = I$) y Λ (lambda mayúscula) diagonal (con ceros fuera de la diagonal principal):

$$B = Q\Lambda Q'$$

Cada uno de los q_i vectores fila de Q es un *autovector*, y cada elemento λ_i (lambda) de Λ es el correspondiente *autovalor*. Existen distintos procedimientos para hallar los valores de cada uno de los q_i vectores fila de Q y, a partir de éstos, los autovalores λ_i correspondientes de Λ (véase Borg y Groenen, 1997, páginas 128-131 para un ejemplo utilizando el *método potencial*). Dado que las matrices de productos escalares son simétricas y no poseen autovalores negativos, podemos reescribir la descomposición anterior de B como:

$$B = (Q\Lambda^{1/2})(Q\Lambda^{1/2})'$$

Si hacemos $(Q\Lambda^{1/2}) = X$, obtenemos la igualdad que nos interesa: $B = XX'$.

El modelo de escalamiento métrico se aplica únicamente a datos medidos en escala de intervalo o razón. Toma como entrada, como hemos dicho, una matriz de proximidades entre n objetos y nos proporciona como salida las coordenadas de los n objetos en r dimensiones del espacio.

Modelo de escalamiento no métrico

Mientras que el modelo de MDS métrico plantea una relación lineal entre las proximidades de entrada y las distancias derivadas por el modelo, el modelo de escalamiento no-métrico plantea una relación de tipo monotónico, creciente entre ambas, es decir, una relación de tipo ordinal. En MDS no-métrico, por tanto, la relación entre proximidades y distancias es únicamente del tipo:

$$\text{si } \delta_{ij} > \delta_{kl}, \text{ entonces } d_{ij} \geq d_{kl}.$$

El procedimiento de MDS no métrico parte de una matriz de proximidades ordinal o de otro tipo, que es transformada en una matriz de proximidades en rangos, ordenados desde 1 hasta $(n^2 - n)/2$. Esta transformación se lleva a cabo simplemente asignando los rangos a las proximidades en función de su tamaño. A continuación, se calculan unos valores transformados, llamados *disparidades* (d_{ij}) que se ajustan monotónicamente a las proximidades. Generalmente se comienza con una configuración de distancias generada aleatoriamente o mediante algún otro método, y se va ajustando ésta hasta que los rangos de las disparidades coincidan en el sentido monotónico con los rangos de las proximidades.

Modelo de escalamiento de diferencias individuales

El modelo MDS de diferencias individuales, también conocido como **modelo ponderado**, constituye en realidad parte de una familia de procedimientos de análisis, conocidos como *modelos euclídeos generalizados*, que tienen en común el hecho de que utilizan como entrada varias matrices de proximidad (una para cada fuente de datos) y que admiten ponderaciones diferentes de las dimensiones del espacio para cada fuente de datos. Los distintos modelos difieren entre sí en el modo en que esta ponderación se lleva a cabo, en el uso de datos métricos o no-métricos, o en el permitir que las dimensiones sean, además de ponderadas, rotadas también de forma diferente para cada fuente de datos.

Lo interesante de estos modelos es que permiten tratar diferencias entre distintas fuentes de datos, como sujetos, grupos o momentos temporales. Los modelos métrico y no métrico también pueden utilizar como entrada varias matrices de proximidad, pero considerando a cada una de éstas como replicaciones de una misma fuente de datos, de tal modo que las diferencias existentes entre las distintas matrices se tratan como si fuesen errores. Sin embargo, es muy posible que estas diferencias no se deban a errores sino que, por el contrario, sean sistemáticas. Los modelos MDS de diferencias individuales permiten incorporar estas diferencias en la solución del análisis.

El más conocido y utilizado de estos modelos es el **modelo INDSCAL** (*Individual Differences SCALing*). El modelo INDSCAL asume una generalización de la distancia euclídea; la distancia euclídea ponderada:

$$d_{ij,k} = \sqrt{\sum_{a=1}^r w_{ka} (x_{ia} - x_{ja})^2}$$

La distancia euclídea ponderada entre los objetos i y j , para la fuente de datos k , es la suma (para las r dimensiones del espacio) de las diferencias cuadráticas entre las coordenadas de ambos objetos en la dimensión a , multiplicadas por el peso w_{ka} que tiene la dimensión a para la fuente de datos k . Así pues, el modelo plantea que todas las fuentes de datos utilizan un mismo número de dimensiones para caracterizar las relaciones existentes entre los objetos, pero que cada una de ellas pondera de forma diferente estas dimensiones. Así pues, existirá una solución común para todas las fuentes de datos, consistente en una matriz $X(n \times r)$ de coordenadas de los n objetos en las r dimensiones. Esta matriz es equivalente a las matrices coordenadas que se obtienen como soluciones de los modelos de MDS métrico y no métrico. La novedad del modelo INDSCAL es que, además de la matriz X , proporciona otra matriz $W(k \times r)$ que contiene los pesos asignados por cada una de las k fuentes de datos a cada una de las r dimensiones.

La función de los pesos w_{ka} es la de “estirar” o “encoger” una dimensión a determinada en función de la importancia que tiene para la fuente de datos (generalmente el sujeto) k . Cuando el peso asignado a una dimensión es pequeño, ésta se ve “encogida”, de modo que su relevancia para explicar las proximidades encontradas entre los objetos es mucho más reducida que si su peso fuese próximo a 1.

El algoritmo INDSCAL se basa, al igual que el modelo de MDS clásico, en la matriz B de productos escalares. La diferencia estriba en que en el caso de INDSCAL disponemos de k matrices B , una por cada fuente de datos, de modo que la igualdad $B = XX'$ se transforma en $B_k = XW^2_kX'$, donde W^2_k es la matriz de pesos cuadráticos para la fuente de datos k y B_k es la matriz de productos escalares correspondiente a la fuente de datos k . La dificultad del algoritmo INDSCAL radica en que aquí es necesario resolver dos conjuntos de parámetros, los correspondientes a las coordenadas de los objetos, por un lado, y los correspondientes a las matrices de pesos individuales, por el otro. La solución es un procedimiento alternante, donde primero se mantiene fija la matriz X y se estiman los W_k , seguido por una estimación de X manteniendo fijos los W_k . El procedimiento continúa iterativamente hasta que se alcanza la convergencia.

Las matrices $X(n \times r)$ y $W(k \times r)$ representan lo que se denomina, respectivamente, el “espacio de estímulos” y el “espacio de sujetos”, donde la palabra “espacio” viene a indicar que se trata de dos configuraciones diferentes que deben tratarse por separado y que no pueden representarse conjuntamente. El espacio de estímulos se interpreta del mismo modo que el obtenido con otros procedimientos de MDS, con la salvedad de que este espacio no es rotatable. Esta peculiaridad se debe al hecho de que la matriz X de donde proviene este espacio ha sido calculada a partir de la matriz W , y viceversa, de modo que la alteración de las coordenadas en la matriz X conlleva, necesariamente, la alteración consiguiente de los pesos contenidos en la matriz W . No obstante, esto no representa una limitación sino, bien al contrario, una ventaja que posibilita que las dimensiones sean directamente interpretables sin necesidad de orientaciones alternativas.

Por su parte, los pesos contenidos en la matriz W se interpretan como vectores en un espacio de r dimensiones donde, a menor ángulo entre un vector y una dimensión dada, mayor importancia de esa dimensión para la fuente de datos representada por el vector. Elevando al cuadrado el peso w_{ka} correspondiente a la fuente de datos k y la dimensión a obtenemos la proporción de varianza de las proximidades proporcionadas por k que es explicada por a . Del mismo modo, si sumamos todos los pesos cuadráticos para una fuente de datos k obtenemos la proporción de varianza en los datos de k que explica la solución INDSCAL. Finalmente, la ponderación de las coordenadas del espacio común representado por X con los pesos w_k , nos proporcionará el espacio de estímulos *individual* correspondiente a la fuente de datos k .

Modelos de escalamiento para datos de preferencia

Normalmente, el MDS se aplica únicamente a datos de proximidad, bien sea obtenidos directamente, bien sea derivados a partir de datos multivariados. Sin embargo, existen modelos de MDS pensados para otro tipo de datos: los datos de dominancia. Para decirlo en pocas palabras, los datos de dominancia proporcionan información acerca del grado en que existen relaciones de precedencia o jerarquía entre éstos. Existen múltiples formas de recoger datos de dominancia, algunas de ellas increíblemente complejas y alambicadas, pero el ejemplo más habitual y sencillo de datos de dominancia son los datos de preferencia. Para obtener este tipo de datos a partir de una muestra de objetos es necesario únicamente solicitar a los sujetos que ordenen estos objetos en función de su preferencia. Esto nos proporcionará una matriz rectangular de preferencias P de dimensiones $n \times m$ (sujetos x objetos), donde cada elemento p_{ij} de la matriz corresponderá a la preferencia del sujeto i por el objeto j .

Existen dos modelos de MDS muy utilizados con datos de preferencia: el *modelo desdoblado (unfolding)* y el *modelo vectorial*. Ambos hacen distintas asunciones sobre los datos de partida y la representación final de los resultados, pero tienen en común el hecho de que toman como entrada una matriz rectangular de preferencias, y de que proporcionan una solución dimensional conjunta para las dos entidades analizadas: los sujetos y los objetos. Esto representa una gran ventaja de este tipo de modelos frente a los modelos vistos anteriormente. Los modelos MDS métrico y no-métrico sólo permiten representar las relaciones entre un conjunto de objetos (que pueden ser estímulos o sujetos, pero no ambos simultáneamente). Por su parte, el modelo INDSCAL nos proporciona un espacio de estímulos, por un lado, y un espacio de sujetos, por el otro, pero ambos espacios son independientes. Sin embargo, los modelos de análisis de preferencia sitúan en un mismo espacio a objetos y sujetos, de modo que pueden estudiarse las relaciones existentes entre los sujetos, las existentes entre los objetos, y las existentes entre objetos y sujetos.

Para facilitar la convergencia de la solución, la mayoría de los algoritmos de MDS para datos de preferencia asumen que los datos son *condicionales por fila*. Esto quiere decir que las proximidades son comparables entre sí sólo dentro de una misma fila, pero no entre filas diferentes. Esta *condicionalidad* permite la existencia de criterios de preferencia diferentes para cada fila (es decir, para cada sujeto).

Modelo de escalamiento desdoblado (unfolding)

El modelo de MDS desdoblado parte de la suposición de que la matriz de preferencias de $P(n \times m)$ es parte de una matriz de proximidades $\Delta((n + m) \times (n + m))$ entre dos conjuntos de objetos, que podemos llamar A (sujetos) y B (objetos).

Esto significa que la matriz de proximidades está incompleta, puesto que tenemos las proximidades entre los elementos de A y B , pero no tenemos las proximidades de los elementos de A entre sí, ni las proximidades de los elementos de B entre sí. Esto implica que el modelo desdoblado es algo más inestable que los otros modelos de MDS. No obstante, dado que el MDS es muy robusto frente a la ausencia de datos, incluso en estas condiciones es posible obtener soluciones con un buen ajuste. El modelo asume, como hemos dicho, que cada preferencia P_{ij} del sujeto i por el objeto j se interpreta como una medida de proximidad entre ambos. Si el sujeto i muestra gran preferencia por el objeto j le asignará la preferencia 1 y, por tanto, ambos deberían encontrarse a poca distancia uno de otro. Por el contrario, si el sujeto i muestra muy poca preferencia por el objeto j , le asignará un rango próximo a m y, por tanto, ambos deberían encontrarse a mucha distancia. Formalmente, la relación entre proximidades (δ_{ij}), preferencias (p_{ij}) y distancias (d_{ij}) es la siguiente:

$$\delta_{ij} = f(p_{ij}) = d_{ij}^2$$

donde f puede ser una función lineal (caso métrico) o monotónica (caso no-métrico), y donde las distancias cuadráticas d_{ij}^2 se interpretan como:

$$d_{ij}^2 = \sum_{a=1}^r (y_{ia} - x_{ja})^2$$

siendo y_{ia} la coordenada del sujeto i en la dimensión a , y x_{ja} , la coordenada del objeto j en la misma dimensión. Esta solución implica que tanto sujetos como objetos aparecen como puntos en un mismo espacio, y que las preferencias de un sujeto deberían estar en correspondencia con la distancia a la que se hallen los objetos del punto que representa al sujeto, de tal modo que cuanto más preferido sea un objeto, más próximo debería encontrarse a ese punto. Interpretado así, el punto que representa al sujeto correspondería al objeto “ideal”, o de máxima preferencia. Por esta razón también se conoce al modelo desdoblado como modelo del “punto ideal”.

Modelo de escalamiento vectorial

El modelo de MDS vectorial se diferencia del modelo desdoblado o del “punto ideal”, en que las filas de la matriz de preferencias (es decir, los sujetos) no se representan mediante puntos, sino mediante vectores de longitud unidad. Lo que pretende el modelo vectorial es encontrar una combinación lineal de los valores de coordenadas de los objetos, de modo que sus proyecciones sobre el vector que representa a un sujeto se correspondan lo más estrechamente posible con las preferencias manifestadas por ese sujeto. Para obtener la solución del modelo vectorial, generalmente se recurre a la descomposición en valores singulares de la matriz P de preferencias:

$$P = K\Lambda L'$$

Las primeras r columnas de $K\Lambda$ y de L nos proporcionarían, respectivamente, la solución en r dimensiones para las matrices de coordenadas de los puntos correspondientes a los objetos, $X(m \times r)$ y de los vectores correspondientes a los sujetos $Y(n \times r)$.

Los modelos de escalamiento parten de una función de representación de las proximidades estimadas

INTERPRETACIÓN DE LOS RESULTADOS OBTENIDOS

La forma más habitual de interpretar las soluciones MDS es la interpretación dimensional. Esta forma supone ordenar los objetos y/o los sujetos a lo largo de continuos (dimensiones) que se interpretan como escalas de medida de alguna característica o atributo. Esto es lo que se entiende generalmente como “escalamiento”. Estas escalas no siempre coinciden en orientación con las dimensiones originalmente proporcionadas por el MDS. Esto se debe a que, en términos de distancias entre objetos, la orientación de los ejes es arbitraria. Por tanto, si alguna orientación alternativa de los mismos facilita la interpretación, podemos rotar la solución a los nuevos ejes (excepto en INDSCAL) y utilizar éstos para interpretar los resultados.

Pero, además de la interpretación dimensional, existen otras muchas formas de interpretar las soluciones proporcionadas por el MDS. En realidad, el modo en que interpretemos la solución dependerá fundamentalmente de los intereses de nuestra investigación (clasificatorios, exploratorios, confirmatorios o teóricos), de modo que podemos utilizar simultáneamente distintos criterios, o combinarlos, con el fin de entender más cabalmente la estructura de los datos.

Veremos ahora algunas de las formas más usuales de interpretación: la interpretación dimensional, la interpretación por agrupamientos de objetos y la interpretación por regiones, propia esta última del MDS confirmatorio. Veremos también, para cada una de ellas, las estrategias y técnicas auxiliares que pueden utilizarse para facilitar nuestro trabajo.

Interpretación dimensional

La interpretación dimensional busca continuos o vectores a lo largo de los cuales interpretar las posiciones de los objetos. Una forma directa de interpretar la solución MDS en forma dimensional consiste en utilizar datos externos (Arce, 1993).

Para llevar a cabo este procedimiento debemos obtener medidas de los objetos en una serie de atributos, y utilizar cada una de estas medidas como variables dependientes en un análisis de regresión múltiple, mientras que como variables independientes utilizaremos las coordenadas del los objetos en la matriz X. Si alguno de los atributos puede ser expresado como una combinación lineal de una o más de las coordenadas de los objetos, entonces ese atributo está relacionado con la solución proporcionada por el análisis. En el caso de que el atributo venga explicado por una sola dimensión, podremos interpretar ésta en función de aquél.

Existen programas desarrollados específicamente para ajustar un vector de atributos a un espacio de objetos, de modo que los valores de los atributos puedan expresarse como una combinación lineal de las coordenadas de los objetos, de un modo semejante a lo que ocurre con los vectores de los sujetos y sus puntuaciones de preferencia en el modelo vectorial de MDS. Además, algunos programas de MDS también permiten imponer la restricción de que las coordenadas de los objetos en la solución final sean una combinación lineal de variables externas.

Interpretación por agrupamientos

En ocasiones puede ocurrir que efectuamos un análisis MDS sobre un conjunto de objetos con fines clasificatorios. Es decir que, aunque puedan interesarnos los criterios en función de los cuales puedan expresarse las proximidades existentes entre los objetos, también puede interesarnos ver si existen agrupamientos de objetos que sean muy similares entre sí, y diferentes del resto. Imaginemos que pedimos a una muestra de sujetos que evalúen una serie de productos de consumo. Podríamos analizar las similaridades entre estos productos mediante MDS, pero también podría interesarnos ver si existen agrupamientos de sujetos en función de sus hábitos de consumo. En este caso, analizaríamos las similaridades entre sujetos mediante MDS. Aquellos sujetos con hábitos de consumo muy similares se encontrarán muy próximos entre sí, y aquellos grupos de sujetos con hábitos muy diferentes se encontrarán alejados entre sí. Esto nos permite identificar a qué segmentos de la población se deben dirigir determinados productos.

Una técnica utilizada habitualmente cuando queremos llevar a cabo agrupamientos es el análisis de conglomerados. A diferencia del MDS, que proporciona soluciones continuas, el análisis de conglomerados proporciona soluciones discretas y (generalmente) jerárquicas. Utilizando ambas técnicas en conjunción resultará más sencillo identificar agrupamientos de objetos. Un interesante muestrario de técnicas y su uso combinado puede verse en Arabie, Carroll y DeSarbo (1987).

Interpretación por regiones

La interpretación por regiones se utiliza generalmente con MDS confirmatorio, una variante de MDS mucho menos conocida que la exploratoria. Para llevar a cabo este tipo de MDS no hace falta un programa especial, sino un diseño especial de la tarea que deben realizar los sujetos. Se presentan a los sujetos estímulos que presentan una combinación conocida de atributos que se supone que están en la base del fenómeno estudiado y se les pide que los clasifiquen (por preferencias, similaridad, etc.). La solución MDS deberá reflejar los atributos que contienen los estímulos, situando en una misma región (llamada isotónica) aquellos estímulos que contienen el mismo atributo.

APLICACIONES DEL MDS Y SU RELACIÓN CON OTRAS TÉCNICAS DE ANÁLISIS DE DATOS

Podemos decir que el MDS es una técnica de amplio espectro. Existen multitud de modelos diferentes, aplicables a multitud de datos diferentes y a distintas escalas de medida. Asimismo, el MDS puede emplearse con diversos fines, ya sean éstos exploratorios o confirmatorios. En este estado de cosas, es de esperar que el MDS entre en relaciones de competencia (o complementariedad, según el caso) con otras técnicas de análisis multivariante. Por otro lado, esta amplitud de espectro implica que su campo de aplicación es igualmente amplio. En efecto, como ya hemos comentado, aunque el MDS nació dentro de la investigación psicológica, en la actualidad se aplica en multitud de áreas de conocimiento, como técnica de clasificación, de reducción de datos, etc. Este apartado pretende proporcionar alguna información acerca del uso del MDS en conjunción con otras técnicas de análisis, de su relación competitiva con otras, así como de su utilidad en contextos aplicados, presentando ejemplos de uso del MDS en investigación real.

Comenzaremos por el uso de MDS conjuntamente con otras técnicas de análisis estadístico. Una de las facetas más importantes donde esta combinación de técnicas es especialmente útil es la interpretación de la configuración obtenida. Este aspecto ya lo hemos tratado anteriormente al hablar de la utilidad de la regresión múltiple para interpretar las dimensiones encontradas, o del análisis de conglomerados para encontrar agrupamientos de estímulos en el espacio multidimensional. Además de estos procedimientos clásicos, también existen procedimientos derivados expresamente para MDS, como el ajuste de propiedades de los estímulos, que permite incluir éstas como vectores en el espacio multidimensional. Otras técnicas, como el análisis factorial, también pueden utilizarse previamente al empleo de MDS con datos de perfil, con el fin de extraer lo esencial de los datos, eliminando la información redundante presente en las variables originales, tal y como hemos comentado al hablar del cálculo de la distancia euclídea a partir de datos de perfil.

No obstante, el papel de las técnicas de reducción de datos, como el propio análisis factorial, o como el análisis de correspondencias, frente al MDS, es más bien competitivo que complementario. La elección de la técnica en tales casos, vendrá determinada por los objetivos del trabajo o por las preferencias del investigador. No obstante, es necesario señalar que, dado que el MDS es una técnica de análisis menos restrictiva en términos de supuestos (linealidad, etc.) y de datos (tamaño muestral, datos métricos o no-métricos, etc.) que otras técnicas de reducción de datos, su aplicabilidad es generalmente mayor, y sus soluciones, más parsimoniosas que las generadas por técnicas como el análisis factorial.

Existen también aplicaciones del MDS en la investigación. Una de las características más interesantes del MDS es la presentación de las soluciones en forma de mapa espacial. Por ello, una de las aplicaciones en la que ha sido empleado es en el estudio de los mapas cognitivos. Un mapa cognitivo es la representación psicológica de un entorno real que permite, por ejemplo, que no nos perdamos paseando por nuestra ciudad. El estudio de las estimaciones de distancia entre puntos geográficos dentro de una ciudad mediante MDS nos permitirá conocer el modo en que los sujetos almacenan la información espacial, así como los sesgos y distorsiones sistemáticas que introducen en sus estimaciones.

El MDS puede ser utilizado también para estudiar múltiples fenómenos sociales, como la conducta de voto y la participación política de las personas, o las diferentes percepciones que tienen de sí mismos y de otros habitantes de distintos países o de distintas comunidades autónomas dentro de un mismo país. A un nivel más psicológico, el MDS puede ser aplicado también al estudio de fenómenos del comportamiento humano, como la conducta agresiva.

La capacidad del MDS para abordar objetos de estudio complejos e influenciados por múltiples variables lo hace especialmente interesante para el estudio de la percepción del medio ambiente y la evaluación psicológica de entornos. Las aplicaciones de la técnica en este contexto de investigación abarcan la evaluación de entornos como edificios, ciudades o paisajes.

ESCALAMIENTO ÓPTIMO Y MULTIDIMENSIONAL EN SPSS

SPSS Y EL ESCALAMIENTO ÓPTIMO

SPSS incorpora procedimientos que implementan las técnicas de escalamiento óptimo. La combinación del nivel de medida de las variables y el número de conjuntos de variables seleccionados para el estudio determinan el procedimiento de escalamiento óptimo mediante mínimos cuadrados alternantes que se realizará. Ya sabemos que el nivel de medida puede ser nominal, ordinal y numérico. Por su parte, el número de conjuntos de variables especifica cuántos grupos de variables se van a comparar con otros grupos de variables. Cuando todas las variables son nominales múltiples, todas las variables del análisis tienen cuantificaciones en varias categorías. Si alguna variable no es nominal múltiple, se entiende que una o más variables en el análisis se escalan a un nivel diferente del nominal múltiple. Otros niveles de escala posibles son nominal simple, ordinal y numérica discreta.

La combinación de opciones para *Nivel de medida* y *Número de conjuntos de variables* proporcionarán un análisis de homogeneidad (correspondencias múltiples), un análisis de componentes principales categóricos o un análisis de correlación canónica no lineal. Las opciones para cada procedimiento en la pantalla de entrada son:

Análisis de correspondencias múltiples u homogeneidad (HOMALS): Seleccione *Todas las variables son nominales múltiples* y *Un conjunto*.

Análisis de componentes principales categóricas (CATPCA): Seleccione *Alguna variable no es nominal múltiple* y *Un conjunto*.

Análisis de correlación canónica no lineal (OVERALS): Seleccione *Múltiples conjuntos*.

Análisis de componentes principales categóricas con SPSS

El análisis estándar de componentes principales asume que todas las variables del análisis se miden a escala numérica, y que las relaciones entre los pares de las variables son lineales. El análisis de componentes principales categóricas extiende esta metodología para permitir la ejecución del análisis de componentes principales en cualquier mezcla de variables nominales, ordinales y numéricas. El objetivo del análisis de componentes principales categóricas sigue siendo contabilizar la mayor variación posible en los datos, dada la dimensionalidad especificada del análisis. Para las variables nominales y ordinales del análisis, el programa calcula las puntuaciones óptimas para las categorías. En la configuración usual de datos de SPSS, las filas son individuos, las columnas son las medidas para los ítems, y las puntuaciones a través de las filas son las puntuaciones de las preferencias. No obstante, utilizando el procedimiento TRANSPOSE de SPSS, se pueden transponer los datos.

Si todas las variables se declaran numéricas, el análisis no lineal de componentes principales (CATPCA) equivale al análisis de componentes principales estándar utilizando el análisis factorial. Si todas las variables se declaran nominales múltiples, el análisis no lineal de componentes principales equivale al análisis de correspondencias múltiples (HOMALS) ejecutado sobre las mismas variables. Por tanto, CATPCA es un tipo de HOMALS con algunas variables ordinales o numéricas.

Como ejemplo, el análisis de componentes principales categórico se puede utilizar para representar gráficamente la relación entre la categoría laboral, la división laboral, la provincia, el número de desplazamientos (alto, medio y bajo) y la satisfacción laboral. Observará que con dos dimensiones se puede explicar una gran cantidad de varianza. La primera dimensión podría separar la categoría laboral de la provincia, mientras que la segunda dimensión podría separar la división laboral del número de desplazamientos. También podrá observar que la alta satisfacción laboral está relacionada con un número medio de desplazamientos.

Como ejemplo adicional podría considerarse la identificación de grupos posibles de sistemas sociales, considerando cinco variables que describen la intensidad de interacción social (*intensid*) clasificada en 4 categorías (ligera, baja, moderada y alta), los sentimiento de pertenencia a un grupo (*pertenen*) clasificada en 4 categorías (ninguno, ligero, variable y alto), la proximidad física de los miembros (*proximid*), clasificada en 2 categorías (distante y cercano), la formalidad de la relación entre los miembros (*formalid*), clasificada en 3 categorías (sin relación, formal e informal) y la frecuencia de interacción entre sus miembros (*frecuenc*) clasificada en 4 categorías (ligera, no recurrente, no frecuente y frecuente).

En cuanto a estadísticos y gráficos, se obtienen: Frecuencias, valores perdidos, nivel de escalamiento óptimo, moda, varianza explicada por las coordenadas del centroide, las coordenadas de vector, total por variable y total por dimensión, saturaciones en los componentes para las variables cuantificadas por los vectores, cuantificaciones y coordenadas de categoría, historial de iteraciones, correlaciones entre las variables transformadas y los autovalores de la matriz de correlaciones, correlaciones entre las variables originales y los autovalores de la matriz de correlaciones, puntuaciones de objetos, gráficos de categorías, gráficos de categorías conjuntas, gráficos de transformación, gráficos de residuos, gráficos de centroides proyectados, gráficos de objetos, diagramas de dispersión biespaciales, diagramas de dispersión triespaciales y gráficos de las saturaciones en los componentes.

En cuanto a los datos, los valores de las variables de cadena se convierten en enteros positivos por orden alfabético ascendente. Los valores perdidos definidos por el usuario, los valores perdidos del sistema y los valores menores que 1 se consideran valores perdidos. Se puede añadir una constante o recodificar las variables con valores inferiores a 1 para evitar que se pierdan los mismos. Los datos deben contener al menos tres casos válidos. El análisis se basa en datos enteros positivos. La opción de discretización categorizará de forma automática una variable con valores fraccionarios, agrupando sus valores en categorías con una distribución casi "normal" y convertirá de forma automática los valores de las variables de cadena en enteros positivos. Se pueden especificar otros esquemas de discretización.

Para realizar un análisis de componentes principales categórico, elija en los menús *Analizar → Reducción de datos → Escalamiento óptimo* (Figura 10-1). Previamente es necesario cargar en memoria el fichero de nombre SOCIAL mediante *Archivo → Abrir → Datos*. Este fichero contiene datos sobre las variables sociológicas definidas anteriormente *intensid, frecuenc, perten, proximid y formalid*.

En el cuadro de diálogo *Escalamiento óptimo* de la Figura 10-2, seleccione *Alguna variable no es nominal múltiple*. A continuación seleccione *Un conjunto*, pulse en *Definir*, y en la Figura 10-3 seleccione dos o más variables para el análisis y especifique el número de dimensiones en la solución (campo *Dimensiones en la solución*). Defina la escala y la ponderación para las variables con el botón *Definir escala y ponderación* (Figura 10-4). Si lo desea, tiene la posibilidad de seleccionar una o más variables para proporcionar etiquetas de punto en los gráficos de las puntuaciones de objeto (campo *Variables de etiquetado*). Cada variable genera un gráfico diferente, con los puntos etiquetados mediante los valores de dicha variable. En el campo *Variables suplementarias* se introducen las variables que no se utilizan para hallar la solución de los componentes principales, pero que posteriormente se ajustan a la solución encontrada. Debe seleccionar la escala de medida (nivel para escalamiento óptimo) para todas las variables suplementarias en el análisis mediante el botón *Definir escala*.

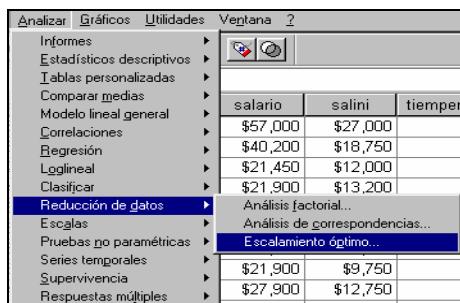


Figura 10-1

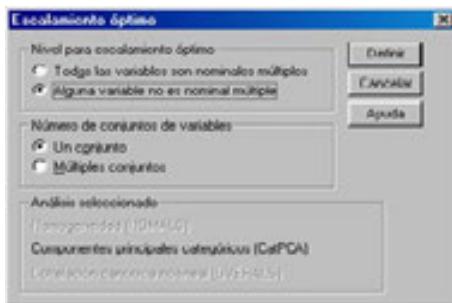


Figura 10-2

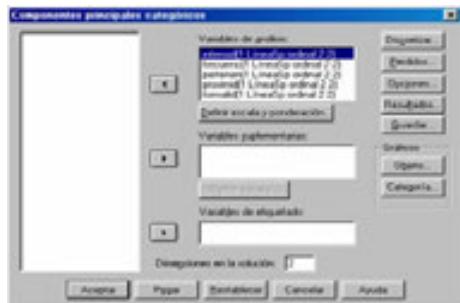


Figura 10-3

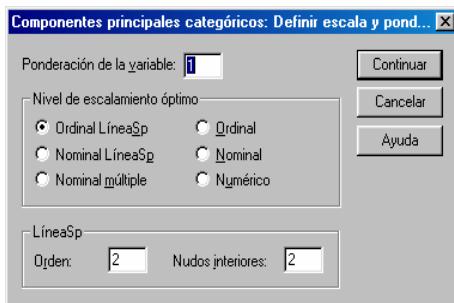


Figura 10-4

Se debe definir el nivel de escalamiento óptimo de las variables del análisis y de las suplementarias (Figura 10-4). Por defecto, se escalan como *línneasSp* (*ordinales*) monotónicas de segundo orden con dos nudos interiores. Asimismo, se puede definir una ponderación para cada variable del análisis mediante la casilla *Ponderación de la variable*. El valor especificado debe ser un entero positivo. El valor por defecto es 1. Las posibles opciones para seleccionar el nivel de escalamiento que se utilizará para cuantificar cada variable son las siguientes:

LíneaSp ordinal: El orden de las categorías de la variable observada se conserva en la variable escalada ópticamente. Los puntos de categoría estarán sobre una recta (vector) que pasa por el origen. La transformación resultante es un polinomio monotónico por tramos suave del orden seleccionado. Las partes se especifican por el número de nudos interiores definido por el usuario y su posición es determinada por el procedimiento en función del número de nudos interiores.

LíneaSp nominal: La única información de la variable observada que se conserva en la variable escalada ópticamente es la agrupación de los objetos en las categorías. No se conserva el orden de las categorías de la variable observada. Los puntos de categoría estarán sobre una recta (vector) que pasa por el origen. La transformación resultante es un polinomio, posiblemente monotónico, por tramos suave del orden seleccionado. Las partes se especifican por el número de nudos interiores definido por el usuario y su posición es determinada por el procedimiento en función del número de nudos interiores.

Nominal múltiple: La única información de la variable observada que se conserva en la variable escalada óptimamente es la agrupación de los objetos en las categorías. No se conserva el orden de las categorías de la variable observada. Los puntos de categoría estarán en el centroide de los objetos para las categorías particulares. El término múltiple indica que se obtienen diferentes conjuntos de cuantificaciones para cada dimensión.

Ordinal: El orden de las categorías de la variable observada se conserva en la variable escalada óptimamente. Los puntos de categoría estarán sobre una recta (vector) que pasa por el origen. La transformación resultante se ajusta mejor que la transformación de líneaSp ordinal pero la suavidad es menor.

Nominal: La única información de la variable observada que se conserva en la variable escalada óptimamente es la agrupación de los objetos en las categorías. No se conserva el orden de las categorías de la variable observada. Los puntos de categoría estarán sobre una recta (vector) que pasa por el origen. La transformación resultante se ajusta mejor que la transformación de líneaSp nominal pero la suavidad es menor.

Numérico: Las categorías se tratan como que están ordenadas y espaciadas uniformemente (a nivel de intervalo). El orden de las categorías y la equidistancia entre los números de las categorías de la variable observada se conservan en la variable escalada óptimamente. Los puntos de categoría estarán sobre una recta (vector) que pasa por el origen. Cuando todas las variables están a nivel numérico, el análisis es análogo al análisis de componentes principales típico.

El botón *Discretizar* de la Figura 10-3 nos lleva al cuadro de diálogo *Discretización* de la Figura 10-5), que permite seleccionar un método para recodificar las variables. Las variables con valores fraccionarios se agrupan en siete categorías (o en el número de valores diferentes de la variable si dicho número es inferior a siete) con una distribución aproximadamente normal, si no se especifica lo contrario. Las variables de cadena se convierten siempre en enteros positivos mediante la asignación de indicadores de categoría en función del orden alfanumérico ascendente. La discretización de las variables de cadena se aplica a estos enteros resultantes. Por defecto, las variables restantes se dejan inalteradas. A partir de ese momento, se utilizan en el análisis las variables discretizadas. El campo *Método* permite seleccionar entre *Agrupación* (se recodifica en un número especificado de categorías o se recodifica por intervalos), *Asignación de rangos* (la variable se discretiza mediante la asignación de rangos a los casos) y *Multiplicación* (los valores actuales de la variable se tipifican, multiplican por 10, redondean y se les suma una constante de manera que el menor valor discretizado sea 1). Existen las siguientes opciones al discretizar variables por agrupación: *Número de categorías* (especifique un número de categorías y si los valores de la variable deben seguir una distribución aproximadamente normal o uniforme en dichas categorías) e *Intervalos iguales* (las variables se recodifican en las categorías definidas por dichos intervalos de igual tamaño debiendo especificar la longitud de los intervalos).



Figura 10-5

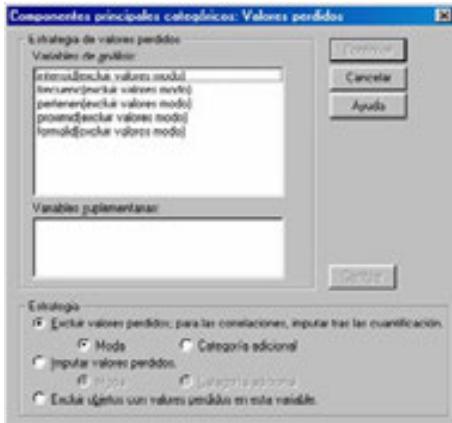


Figura 10-6

El botón *Valores perdidos* de la Figura 10-3 nos lleva al cuadro de diálogo *Valores perdidos* de la Figura 10-6, que permite seleccionar la estrategia para el tratamiento de los valores perdidos en las variables de análisis y las suplementarias. En el campo *Estrategia* seleccione *Excluir los valores perdidos* (tratamiento pasivo), *Imputar los valores perdidos* (tratamiento activo) o *Excluir objetos con valores perdidos* (eliminación por lista).

Si se elige *Excluir valores perdidos*, para las correlaciones, imputar tras la cuantificación. Los objetos con valores perdidos en la variable seleccionada no contribuyen en el análisis de esta variable. Si se especifican correlaciones en el cuadro de diálogo *Resultados*, tras el análisis, los valores perdidos se imputarán con la categoría más frecuente, la moda, de la variable para las correlaciones de la variable original. Para las correlaciones de la variable escalada óptimamente, se puede seleccionar el método de imputación. Seleccione *Moda* para reemplazar los valores perdidos con la moda de la variable escalada óptimamente. Seleccione *Categoría adicional* para reemplazar los valores perdidos con la cuantificación de una categoría adicional. Esto implica que los objetos con un valor perdido en esta variable se consideran que pertenecen a la misma categoría (la adicional).

Si se elige *Imputar valores perdidos*, los objetos con valores perdidos en la variable seleccionada tendrán dichos valores imputados. Se puede seleccionar el método de imputación. Seleccione *Moda* para reemplazar los valores perdidos con la categoría más frecuente. Cuando existen varias modas, se utiliza la que tiene el indicador de categoría más pequeño. Seleccione *Categoría adicional* para reemplazar los valores perdidos con la misma cuantificación de una categoría adicional. Esto implica que los objetos con un valor perdido en esta variable se consideran que pertenecen a la misma categoría (la adicional).

Si se elige *Excluir objetos con valores perdidos en esta variable*, los objetos con valores perdidos en la variable seleccionada se excluyen del análisis. Esta estrategia no está disponible para las variables suplementarias.

El botón *Opciones* de la Figura 10-3 nos lleva al cuadro de diálogo de opciones de la Figura 10-7, que permite seleccionar la configuración inicial, especificar los criterios de iteración y convergencia, seleccionar un método de normalización, elegir el método para etiquetar los gráficos y especificar objetos suplementarios. En el campo *Objetos suplementarios*, especifique el número de caso del objeto que desea convertir en suplementario y después añádalo a la lista. Si se especifica un objeto como suplementario, se ignorarán las ponderaciones de caso para dicho objeto.

En el campo *Método de normalización*, se puede especificar una de las cinco opciones para normalizar las puntuaciones de objeto y las variables. Sólo se puede utilizar un método de normalización en un análisis dado. El método *Principal por variable* optimiza la asociación entre las variables. Las coordenadas de las variables en el espacio de los objetos son las saturaciones en los componentes (las correlaciones con los componentes principales, como son las dimensiones y las puntuaciones de los objetos). Esta opción es útil cuando el interés principal está en la correlación entre las variables. El método *Principal por objeto* optimiza las distancias entre los objetos. Esta opción es útil cuando el interés principal está en las diferencias y similaridades entre los objetos. El método *Simétrico* utiliza esta opción de normalización si el interés principal está en la relación entre objetos y variables. El método *Independiente* utiliza esta opción de normalización si se desea examinar por separado las distancias entre los objetos y las correlaciones entre las variables. El método *Personalizado* permite especificar cualquier valor real en el intervalo cerrado $[-1, 1]$. Un valor 1 es igual al método *Principal por objeto*, un valor 0 es igual al método *Simétrico* y un valor -1 es igual al método *Principal por variable*. Si se especifica un valor mayor que -1 y menor que 1, se puede distribuir el autovalor entre los objetos y las variables. Este método es útil para generar diagramas de dispersión biespaciales y triespaciales a medida.

En el campo *Criterios*, se puede especificar el número máximo de iteraciones que el procedimiento puede realizar durante los cálculos. También se puede seleccionar un valor para el criterio de convergencia. El algoritmo detiene la iteración si la diferencia del ajuste total entre las dos últimas iteraciones es menor que el valor de convergencia o si se ha alcanzado el número máximo de iteraciones.

En el campo *Configuración* se pueden leer datos de un archivo que contenga las coordenadas de una configuración. La primera variable del archivo deberá contener las coordenadas para la primera dimensión, la segunda variable las coordenadas para la segunda dimensión, y así sucesivamente. La opción *Inicial* significa que la configuración del archivo especificado se utilizará como el punto inicial del análisis.

La opción *Fija* significa que la configuración del archivo especificado se utilizará para ajustar las variables. Las variables que se ajustan se deben seleccionar como variables de análisis, pero al ser la configuración fija, se tratan como variables suplementarias (de manera que no es necesario seleccionarlas como variables suplementarias).

El campo *Etiquetar gráficos con* permite especificar si se utilizarán en los gráficos las etiquetas de variable y las etiquetas de valor o los nombres de variable y los valores. También se puede especificar una longitud máxima para las etiquetas.



Figura 10-7



Figura 10-8

El botón *Resultados* de la Figura 10-3 nos lleva al cuadro de diálogo de la Figura 10-8, que permite producir tablas para las puntuaciones de los objetos, las saturaciones en los componentes, el historial de iteraciones, las correlaciones de las variables originales y de las transformadas, la varianza explicada por variable y por dimensión, las cuantificaciones de las categorías para las variables seleccionadas y estadísticos descriptivos para las variables seleccionadas. *Puntuaciones de los objetos* muestra las puntuaciones de los objetos y tiene las siguientes opciones: *Incluir categorías de* (muestra los indicadores de las categorías de las variables de análisis seleccionadas), *Etiquetar puntuaciones de los objetos por* (de la lista de variables especificadas como variables de etiquetado, se puede seleccionar una para etiquetar los objetos). *Saturaciones en componentes* muestra las saturaciones en los componentes para todas las variables que no recibieron niveles de escalamiento nominal múltiple. *Historial de iteraciones* muestra en cada iteración la varianza explicada, la pérdida y el incremento en la varianza explicada. *Correlaciones de variables originales* muestra la matriz de correlaciones de las variables originales y los autovalores de dicha matriz. *Correlaciones de variables transformadas* muestra la matriz de correlaciones de las variables transformadas (mediante escalamiento óptimo) y los autovalores de dicha matriz. *Varianza explicada* muestra la cantidad de varianza explicada por las coordenadas de los centroides, las coordenadas de vectores y total (coordenadas de centroides y de vectores combinadas) por variable y por dimensión. *Cuantificaciones de categorías* muestra las cuantificaciones de las categorías y las coordenadas para cada dimensión de las variables seleccionadas. *Estadísticos descriptivos* muestra las frecuencias, el número de valores perdidos y la moda de las variables seleccionadas.

El botón *Guardar* de la Figura 10-3 nos lleva al cuadro de diálogo de la Figura 10-9, que permite añadir las variables transformadas, las puntuaciones de objeto y las aproximaciones en el archivo de datos de trabajo (*Guardar*) o como nuevas variables en archivos externos, así como guardar los datos discretizados como variables nuevas en un archivo de datos externo (*Guardar en un archivo externo*).



Figura 10-9

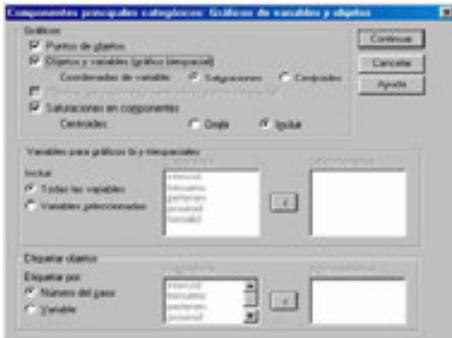


Figura 10-10

El botón *Objeto* del campo *Gráficos* de la Figura 10-3 nos lleva a cuadro de diálogo *Gráficos de variables y objetos* de la Figura 10-10, que permite especificar los tipos de gráficos deseados y las variables para las que se generarán los gráficos. En cuanto al campo *Gráficos*, la casilla *Puntos de objetos* muestra un gráfico de los puntos de objetos. La casilla *Objetos y variables (gráfico biespatial)* muestra un gráfico donde los puntos de objetos se representan con la selección realizada para las coordenadas de las variables: saturaciones en los componentes o centroides de las variables. La casilla *Objetos, saturaciones y centroides (gráfico triespacial)* muestra un gráfico donde los puntos de objetos se representan con los centroides de las variables con un nivel de escalamiento nominal múltiple y las saturaciones en los componentes de las otras variables. La casilla *Saturaciones en componente* muestra un gráfico de las saturaciones en los componentes. Las variables con un nivel de escalamiento nominal múltiple no tienen saturaciones en los componentes, pero se pueden incluir los centroides de dichas variables en el gráfico. En el campo *Variables para gráficos biespaciales y triespaciales* puede utilizar todas las variables para los gráficos de dispersión biespatial y triespatial o seleccionar un subconjunto. En el campo *Etiquetar objetos* se puede elegir que los objetos se etiqueten con las categorías de las variables seleccionadas (se pueden seleccionar entre los valores del indicador de categoría o las etiquetas de valor, en el cuadro de diálogo *Opciones*) o con sus números de caso. Se genera un gráfico por cada variable, si se especifica *Variable*.

El botón *Categorías* del campo *Gráficos* de la Figura 10-3 nos lleva al cuadro de diálogo *Gráficos de categorías* de la Figura 10-11, que permite especificar los tipos de gráficos deseados y las variables para las que se generarán los gráficos. El campo *Gráficos de categorías* permite, para cada variable seleccionada, representar un gráfico de las coordenadas de vector y del centroide.

Para las variables con nivel de escalamiento nominal múltiple, las categorías están sobre los centroides de los objetos para las categorías particulares. Para todos los demás niveles de escalamiento, las categorías están sobre un vector que pasa por el origen. El campo *Gráficos de categorías conjuntas* permite realizar un único gráfico con el centroide y las coordenadas de vector de cada variable seleccionada. El campo *Gráficos de transformación* muestra un gráfico de las cuantificaciones de las categorías óptimas en oposición a los indicadores de las categorías. Se puede especificar el número de dimensiones deseado para las variables con nivel de escalamiento nominal múltiple; se generará un gráfico para cada dimensión. También se puede seleccionar si se muestran los gráficos de los residuos para cada variable seleccionada. En el campo *Proyectar los centroides de*, se puede seleccionar una variable y proyectar sus centroides sobre las variables seleccionadas. Las variables con niveles de escalamiento nominal múltiple no se pueden seleccionar para la proyección. Al solicitar este gráfico, aparece una tabla con las coordenadas de los centroides proyectados.

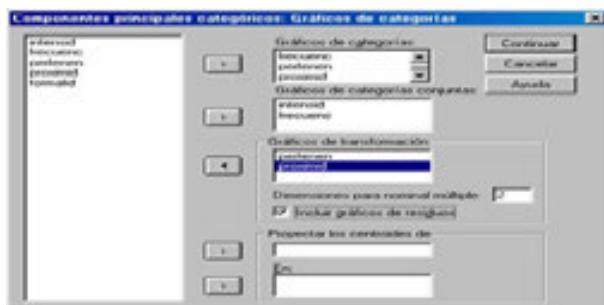


Figura 10-11

Una vez elegidas las especificaciones (que se aceptan con el botón *Continuar*), se pulsa el botón *Aceptar* en la Figura 10-3 para obtener los resultados del análisis según se muestra en la Figura 10-12. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 10-13 a 10-25 se presentan varias salidas tabulares y gráficas del procedimiento.

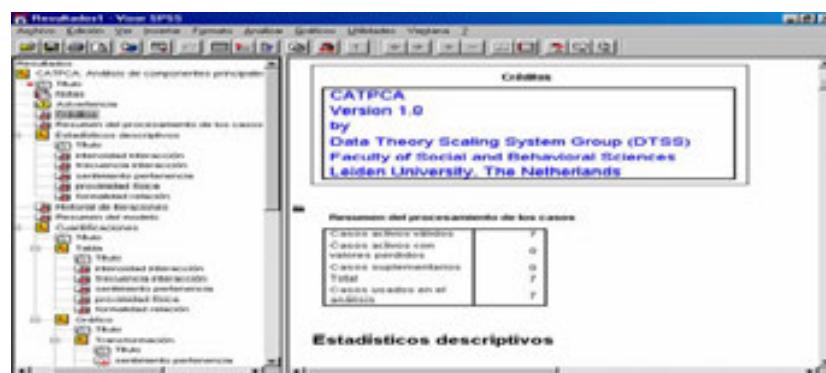


Figura 10-12



Figura 10-13

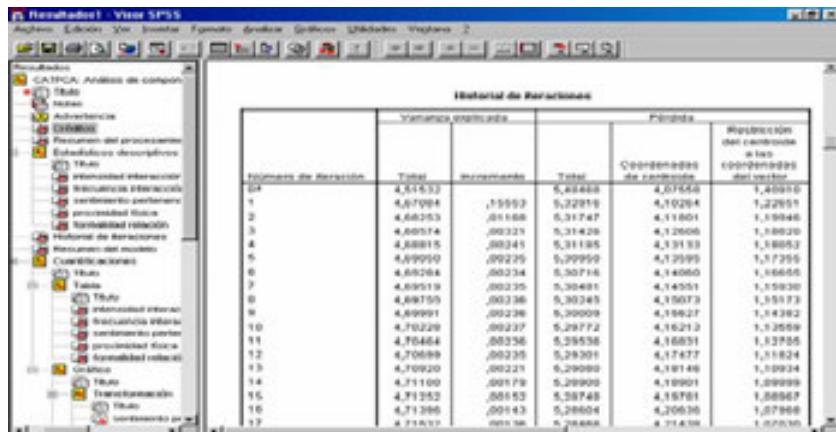


Figura 10-14



Figura 10-15

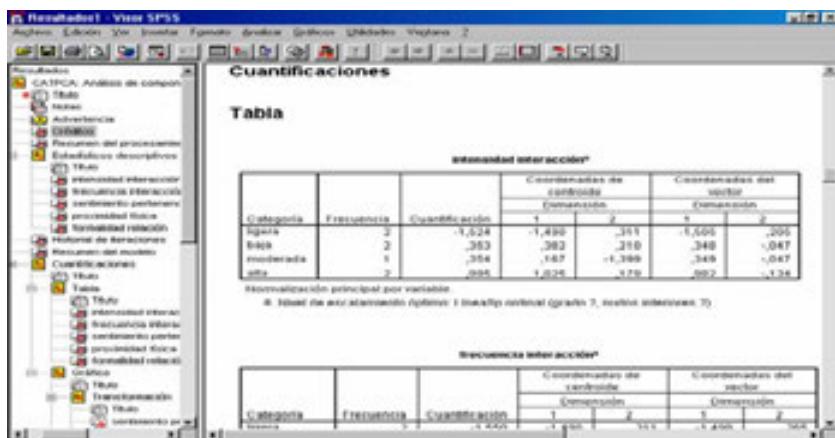


Figura 10-16

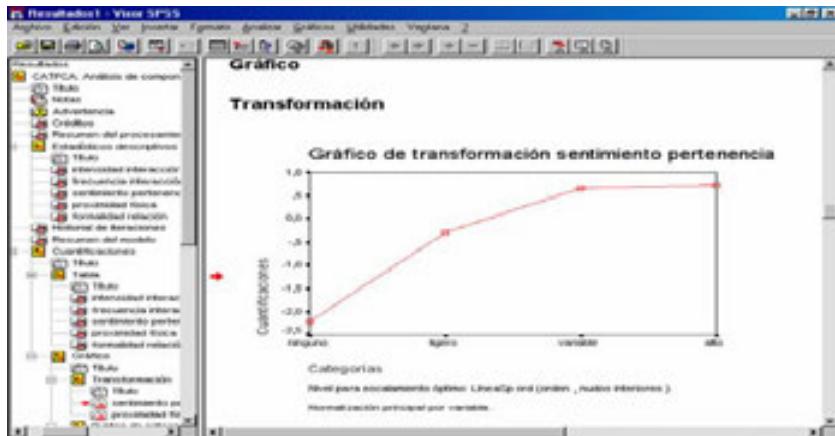


Figura 10-17

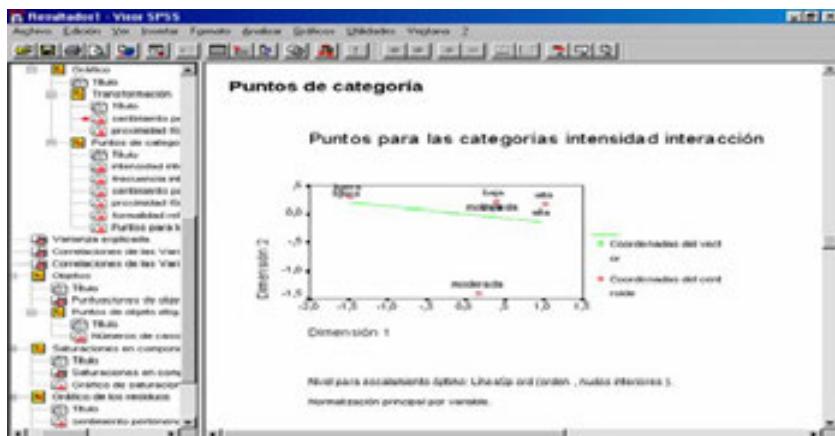


Figura 10-18

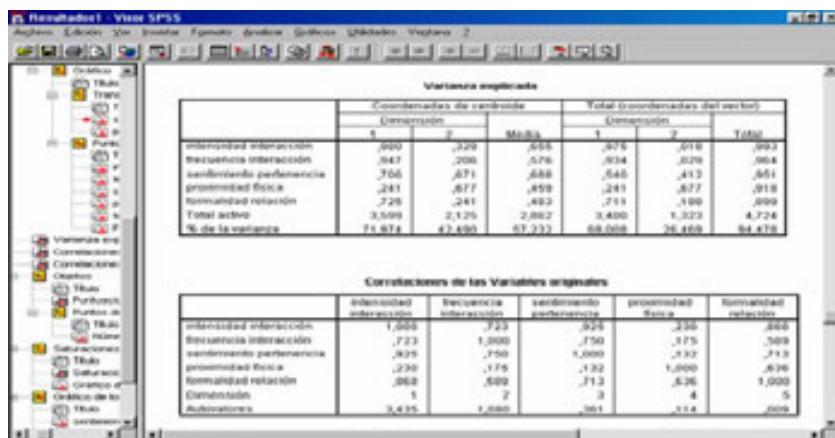


Figura 10-19

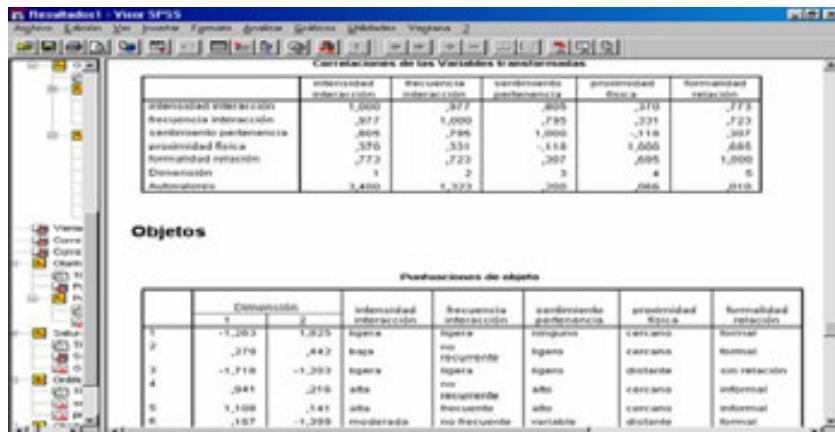


Figura 10-20

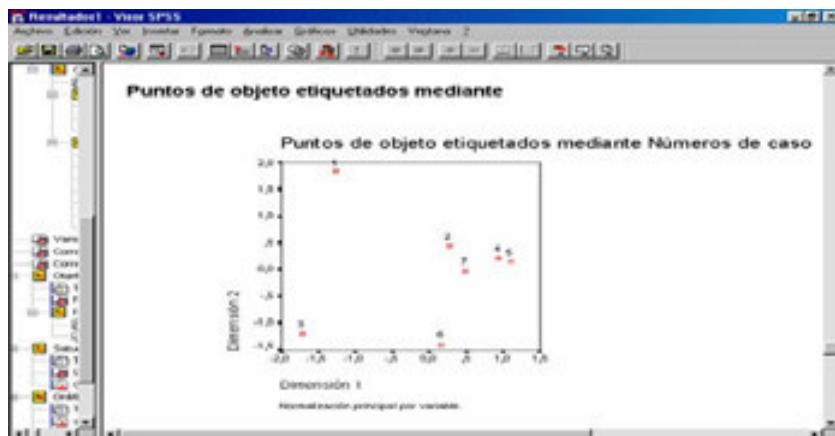


Figura 10-21



Figura 10-22

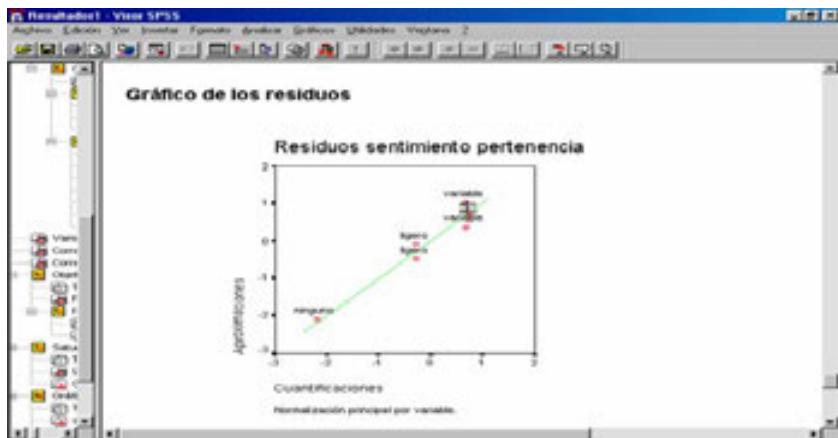


Figura 10-23

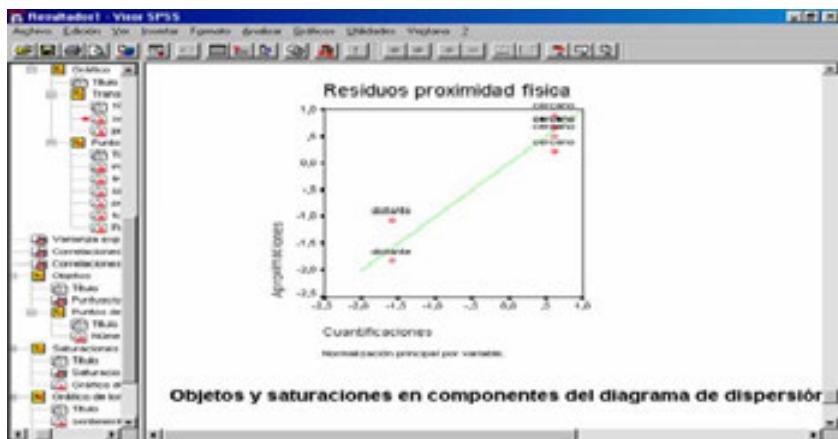


Figura 10-24

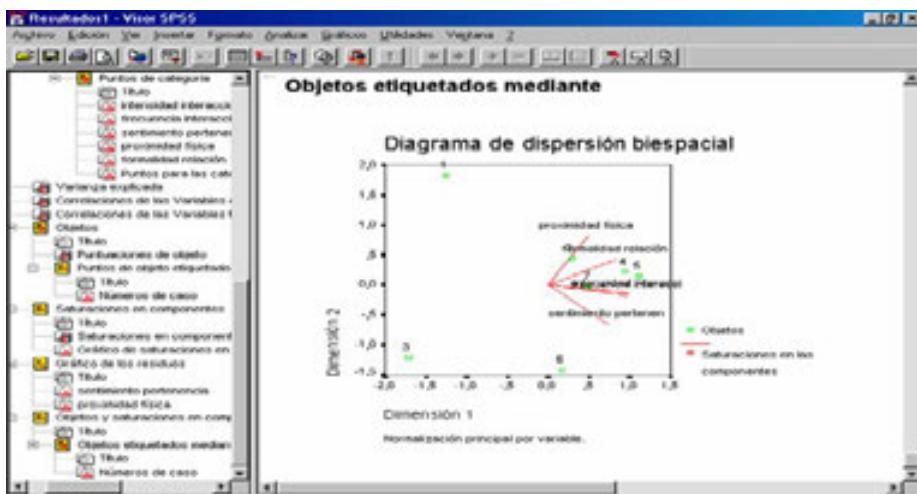


Figura 10-25

Correlación canónica no lineal con SPSS

El procedimiento OVERALS de SPSS, que ejecuta el análisis no lineal de la correlación canónica sobre dos o más grupos de variables es el más general de la familia del escalamiento óptimo.

Como ejemplo, el análisis de correlación canónica categórica mediante escalamiento óptimo se puede utilizar para representar gráficamente la relación entre un conjunto de variables que contienen la categoría laboral (*catlab*) y el nivel educativo (*educ*) y otro conjunto de variables con la clasificación étnica (*minoría*) y el género (*sexo*) de los empleados de una empresa. Podemos encontrar que los años de formación y la clasificación étnica discriminan mejor que las variables restantes. También podemos encontrar que la categoría laboral es la variable que mejor discrimina en la primera dimensión.

En cuanto a estadísticos y gráficos se obtienen: Frecuencias, centroides, historial de iteraciones, puntuaciones de objeto, cuantificaciones de categoría, ponderaciones, saturaciones en las componentes, ajuste simple y múltiple, gráficos de las puntuaciones de objeto, gráficos de las coordenadas de categoría, gráficos de las saturaciones en las componentes, gráficos de los centroides de categoría y gráficos de transformación.

Para realizar un análisis de correlación canónica no lineal, elija en los menús *Analizar → Reducción de datos → Escalamiento óptimo* (Figura 10-26). Previamente es necesario cargar en memoria el fichero de nombre *EMPLEADOS* mediante *Archivo → Abrir → Datos*. Este fichero contiene datos sobre los trabajadores de una empresa con las variables *catlab*, *educ*, *minoría* y *sexo* antes descritas.

En el cuadro de diálogo *Escalamiento óptimo* de la Figura 10-27, seleccione *Todas las variables son nominales múltiples o Alguna variable(s) no es nominal múltiple*. A continuación seleccione *Múltiples conjuntos* y pulse en *Definir*. Defina al menos dos conjuntos de variables. Seleccione la variable o variables que deseé incluir en el primer conjunto (Figura 10-28). Para desplazarse al siguiente conjunto, pulse en *Siguiente* y seleccione las variables que deseé incluir en el segundo conjunto (Figura 10-29). Puede añadir los conjuntos adicionales que deseé. Pulse en *Anterior* para volver al conjunto de variables definido anteriormente. Defina los rangos para las variables con el botón *Definir rango y escala* (Figura 10-30). Si lo desea, tiene la posibilidad de seleccionar una o más variables para proporcionar etiquetas de punto en los gráficos de las puntuaciones de objeto (campo *Etiquetar gráficos de las puntuaciones de objeto con*). Cada variable genera un gráfico diferente, con los puntos etiquetados mediante los valores de dicha variable. Debe definir un rango para cada una de las variables de etiquetado de los gráficos (botón *Definir rango*).

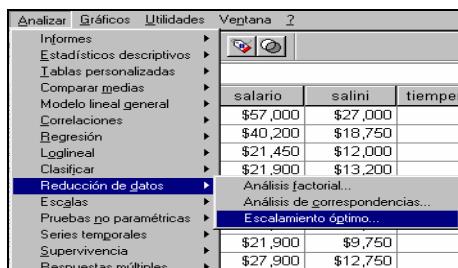


Figura 10-26

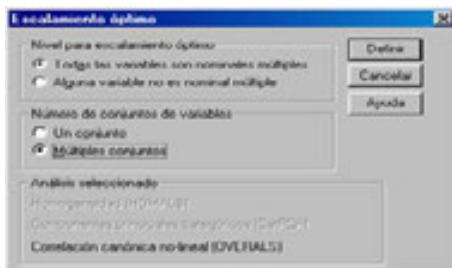


Figura 10-27



Figura 10-28



Figura 10-29

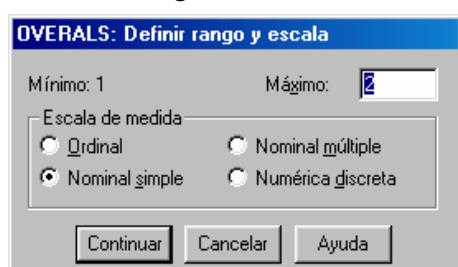


Figura 10-30

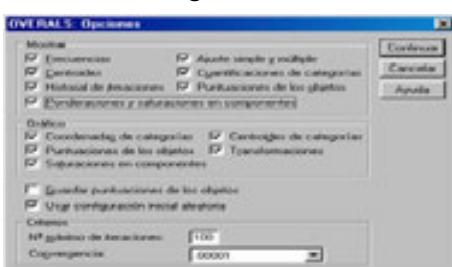


Figura 10-31

El botón *Opciones* (Figura 10-31) permite seleccionar estadísticos y gráficos opcionales, guardar en el archivo de datos de trabajo las puntuaciones de los objetos como nuevas variables y, por último, especificar los criterios de iteración y de convergencia. En el campo *Mostrar* los estadísticos disponibles incluyen las frecuencias marginales (los recuentos), los centroides, el historial de iteraciones, las ponderaciones y las saturaciones en las componentes, las cuantificaciones de las categorías, las puntuaciones de objeto y los estadísticos de ajuste simple y múltiple. En el campo *Gráfico* puede generar gráficos de las coordenadas de las categorías, las puntuaciones de objeto, las saturaciones en las componentes, los centroides de las categorías y las transformaciones. La casilla *Guardar puntuaciones de los objetos* permite guardar las puntuaciones de los objetos como nuevas variables en el archivo de datos de trabajo. Las puntuaciones de objeto se guardan para el número de dimensiones especificadas en el cuadro de diálogo principal. La casilla *Utilizar configuración inicial aleatoria* permite definir una configuración inicial aleatoria en el caso de que todas o algunas de las variables sean nominales simples. Si esta opción no se selecciona, se utiliza una configuración inicial anidada. El campo *Criterios* puede especificar el número máximo de iteraciones que el análisis de correlación canónica no lineal puede realizar durante los cálculos. También puede seleccionar un valor para el criterio de convergencia. El análisis detiene la iteración si la diferencia del ajuste total entre las dos últimas iteraciones es menor que el valor de convergencia o si se ha alcanzado el número máximo de iteraciones.

En cuanto a los datos, utilice enteros para codificar las variables categóricas (nivel de escalamiento nominal u ordinal). Para minimizar los resultados, utilice enteros consecutivos, comenzando por el 1, para codificar cada variable. Las variables escaladas a nivel numérico no deben ser recodificadas en enteros consecutivos. Para minimizar los resultados, en cada variable escalada a nivel numérico, sustraiga el menor valor observado a todos los valores y súmele 1. Los valores fraccionarios se truncarán tras el decimal. Las variables se pueden clasificar en dos o más conjuntos. Las variables del análisis se escalan como nominales múltiples, nominales simples, ordinales o numéricas. El número máximo de dimensiones utilizado en el procedimiento depende del nivel de escalamiento óptimo de las variables. Si todas las variables están especificadas como ordinales, nominales simples o numéricas, el número máximo de dimensiones es el mínimo del número de observaciones menos 1 y el número total de variables. Sin embargo, si sólo se definen dos conjuntos de variables, el número máximo de dimensiones es el número de variables en el conjunto más pequeño. Si algunas variables son nominales múltiples, el número máximo de dimensiones es el número total de categorías nominales múltiples más el número de variables nominales no múltiples menos el número de variables nominales múltiples. Por ejemplo, si el análisis incluye cinco variables, una de las cuales es nominal múltiple con cuatro categorías, el número máximo de dimensiones será $(4 + 4 - 1)$ o 7. Si se especifica un número mayor que el máximo, se utilizará el valor máximo. Una vez elegidas las especificaciones (que se aceptan con el botón *Continuar*), se pulsa el botón *Aceptar* en la Figura 10-29 para obtener los resultados del análisis de correspondencias múltiples según se muestra en la Figura 10-32.

En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 10-33 a 10-40 se presentan varias salidas tabulares y gráficas de entre las múltiples que ofrece el procedimiento.

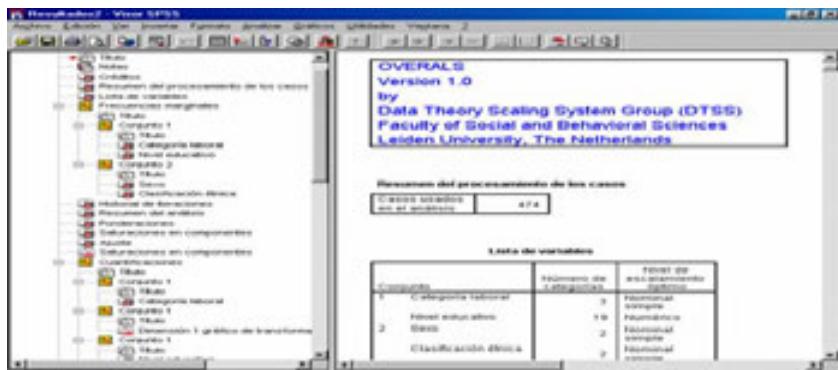


Figura 10-32

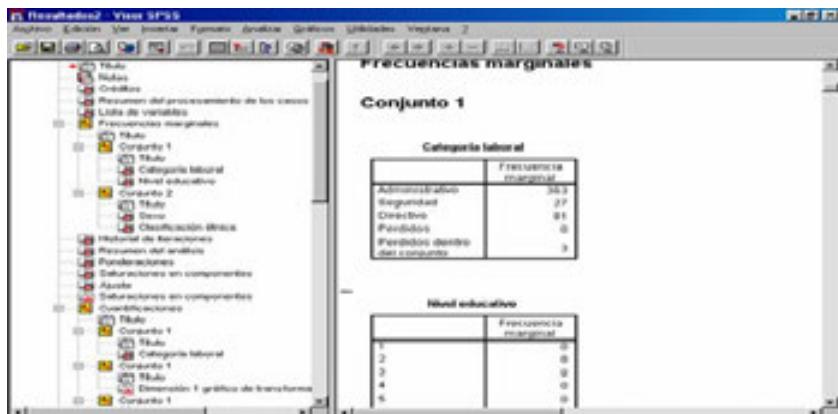


Figura 10-33

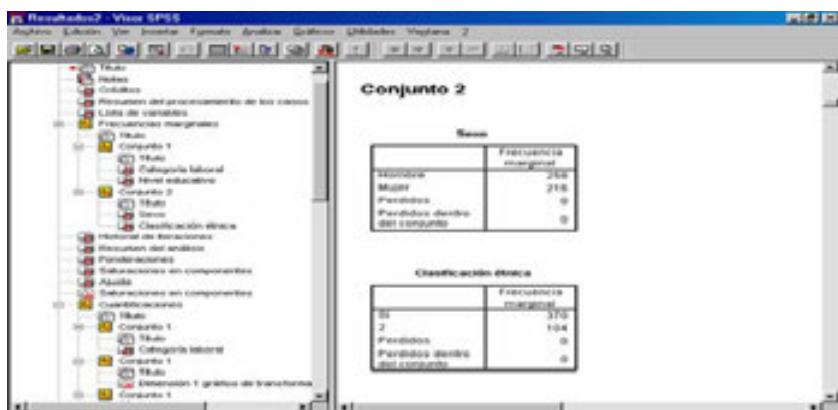


Figura 10-34

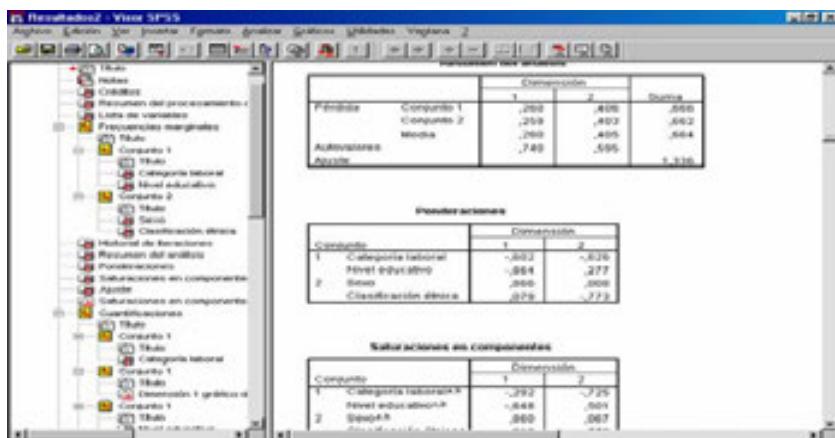


Figura 10-35

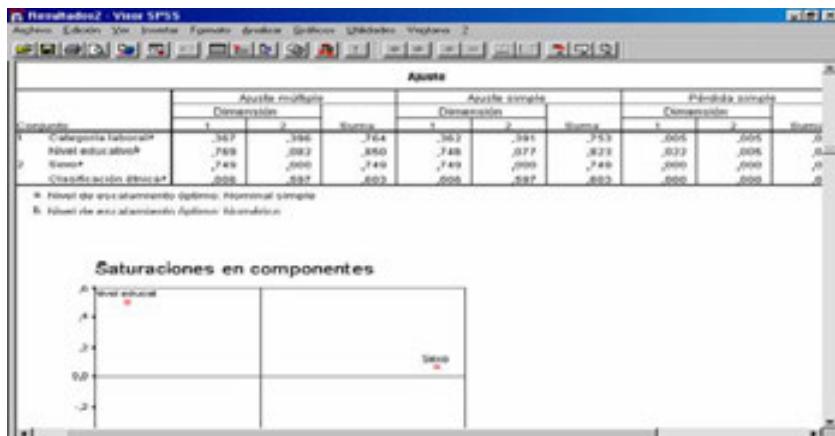


Figura 10-36

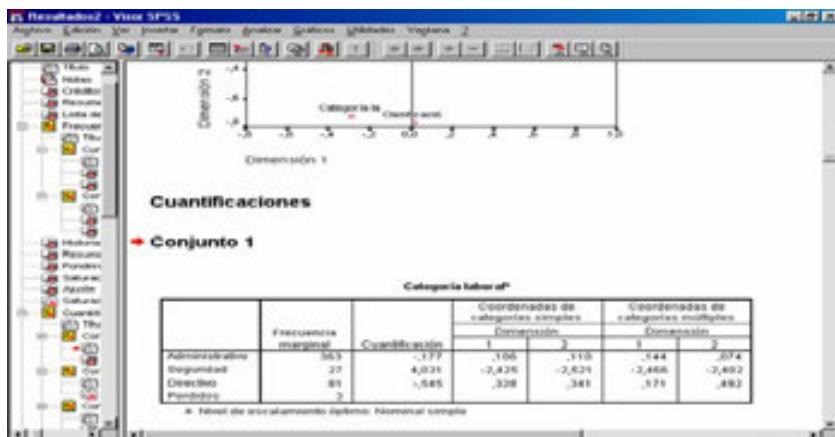


Figura 10-37

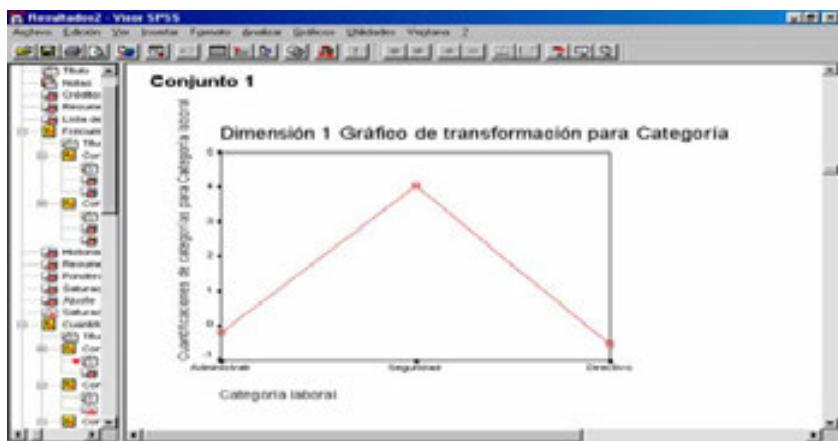


Figura 10-38

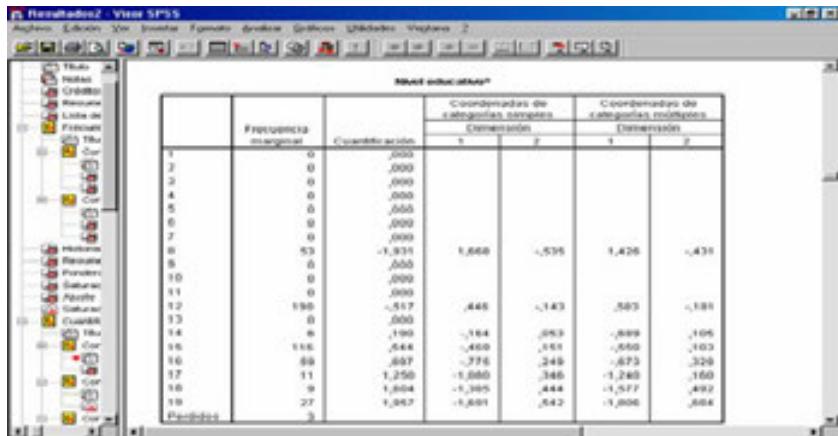


Figura 10-39



Figura 10-40

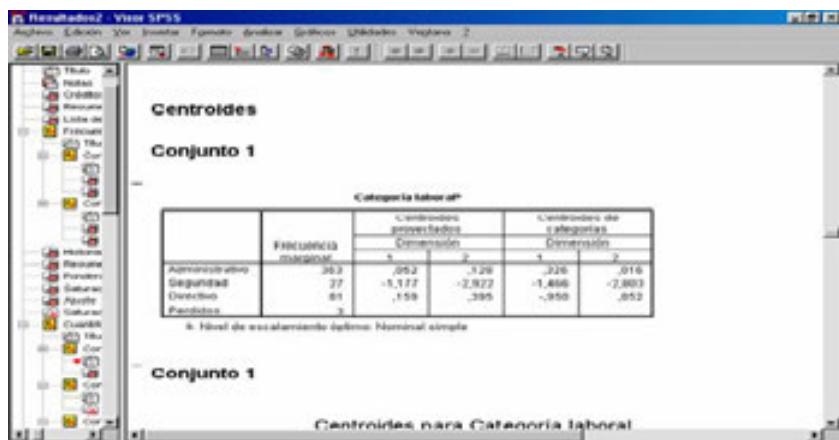


Figura 10-41

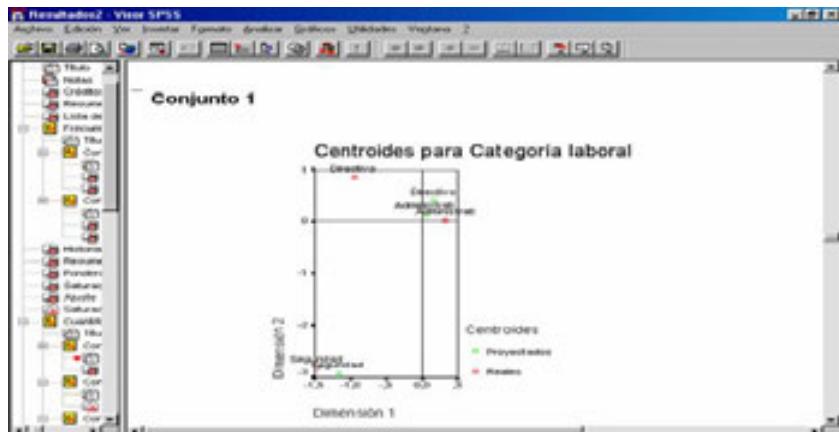


Figura 10-42

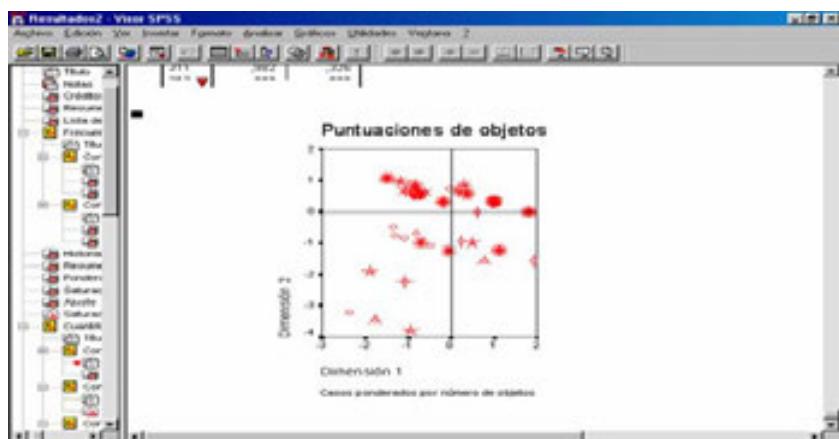


Figura 10-43

SPSS Y EL ESCALAMIENTO MULTIDIMENSIONAL

SPSS incorpora el procedimiento Escalamiento multidimensional (ALSCAL) que trata las técnicas EMD. En cuanto a los datos de entrada, si son de disimilitud, todas ellas deben ser cuantitativas y deben estar medidas en la misma métrica. Si los datos son datos multivariantes, las variables pueden ser datos cuantitativos, binarios o de recuento. El escalamiento de las variables es un tema importante, ya que las diferencias en el escalamiento pueden afectar a la solución. Si las variables tienen grandes diferencias en el escalamiento (por ejemplo, una variable se mide en dólares y otra en años), debe considerar el tipificarlas (esto puede llevarse a cabo automáticamente con el propio procedimiento Escalamiento multidimensional). El procedimiento Escalamiento multidimensional está relativamente libre de supuestos distribucionales. Compruebe que selecciona el nivel de medida adecuado (ordinal, de intervalo, o de razón) en Opciones para asegurar que los resultados son correctos.

Procedimiento ALSCAL

Para realizar un análisis EMD mediante ALSCAL, elija en los menús *Analizar* → *Escala* → *Escalamiento multidimensional* (Figura 10-44). Previamente es necesario cargar en memoria el fichero de nombre DISTAN mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene datos sobre la matriz de distancias entre cinco puntos estratégicos en la distribución comercial. En *Distancias* (Figura 10-45), seleccione *Los datos son distancias* o bien *Crear distancias a partir de datos*. Si los datos son distancias, debe seleccionar al menos cuatro variables numéricas para el análisis y puede pulsar en *Forma* para indicar la forma de la matriz de distancias. Si quiere que SPSS cree las distancias antes de analizarlas, debe seleccionar al menos una variable numérica y puede pulsar en *Medida* para especificar el tipo de medida de distancia que desea. Puede crear matrices distintas para cada categoría de una variable de agrupación (la cual puede ser numérica o de cadena) moviendo esa variable a la lista *Matrices individuales para*.

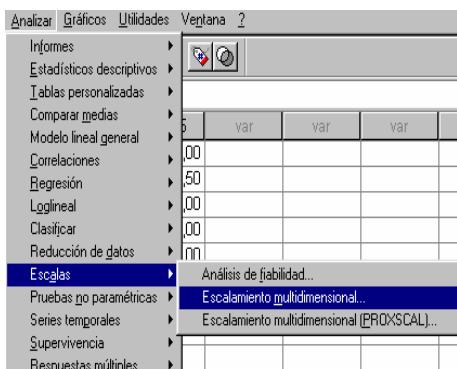


Figura 10-44

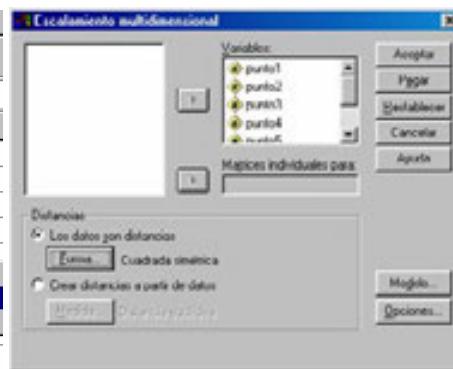


Figura 10-45

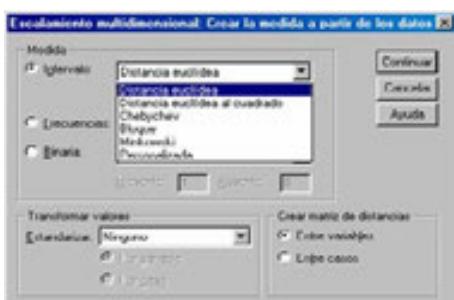


Figura 10-46

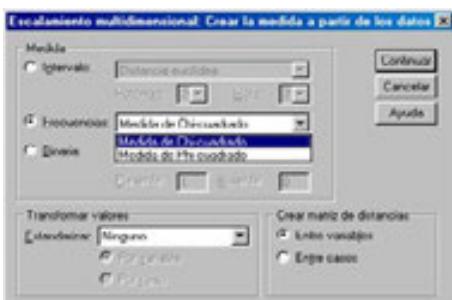


Figura 10-47

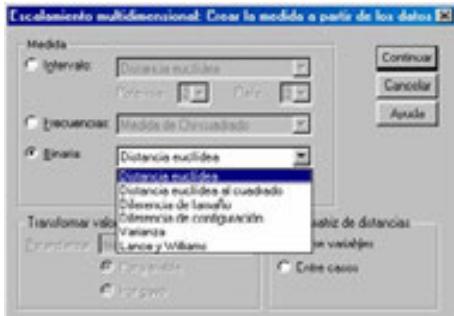


Figura 10-48

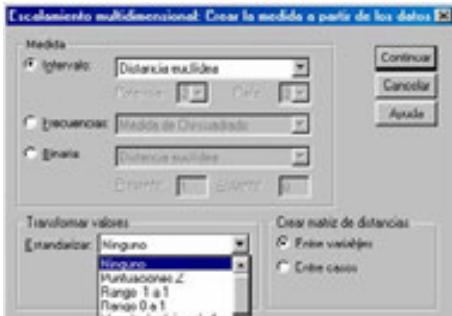


Figura 10-49

El escalamiento multidimensional puede utilizar datos de disimilitud para crear una solución de escalamiento. Si los datos son datos multivariantes (los valores de las variables medidas), debe crear los datos de disimilitud para poder calcular una solución de escalamiento multidimensional. Puede especificar los detalles para la creación de las medidas de disimilitud a partir de los datos utilizando el botón *Crear distancias a partir de datos* de la Figura 10-45. En este caso, el campo *Medida* permite especificar la medida de disimilitud para el análisis.

Seleccione una opción del grupo *Medida* que se corresponda con el tipo de datos y, a continuación, seleccione una de las medidas de la lista desplegable correspondiente a ese tipo de medida. La opción *Intervalo* permite seleccionar entre *Distancia euclídea*, *Distancia euclídea al cuadrado*, *Chebychev*, *Bloque*, *Minkowski* o *Personalizada* (Figura 10-46). La opción *Frecuencia* permite elegir entre *Medida de chi-cuadrado* o *Medida de phi-cuadrado* (Figura 10-47). La opción *Binaria* permite elegir entre *Distancia euclídea*, *Distancia euclídea al cuadrado*, *Diferencia de tamaño*, *Diferencia de configuración*, *Varianza* o *Lance y Williams* (Figura 10-48).

El campo *Crear matriz de distancias* permite elegir la unidad de análisis. Las opciones son *Entre variables* o *Entre casos*. El campo *Transformar valores* permite su estandarización. En determinados casos, como cuando las variables se miden en escalas muy distintas, puede que desee tipificar los valores antes de calcular las proximidades (no es aplicable a datos binarios). Seleccione un método de tipificación de la lista desplegable *Estandarizar* de la Figura 10-49 (si no se requiere ninguna tipificación, seleccione *Ninguna*).

Si el archivo de datos de trabajo representa distancias entre uno o dos conjuntos de objetos, debe elegir el botón *Los datos son distancias* de la Figura 10-45 para especificar la forma de la matriz de datos para obtener los resultados correctos. Elija una opción entre *Cuadrada simétrica*, *Cuadrada asimétrica* o bien *Rectangular* (Figura 10-50). No puede seleccionar *Cuadrada simétrica* si el cuadro de diálogo *Modelo* especifica condicionalidad de filas (Figura 10-51).

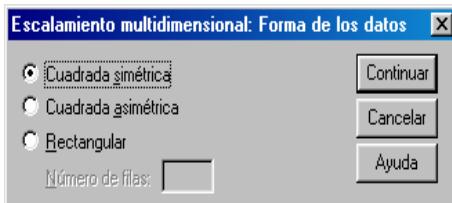


Figura 10-50



Figura 10-51

La estimación correcta de un modelo de escalamiento multidimensional (que se elige haciendo clic el botón *Modelo* de la Figura 10-45 para obtener la Figura 10-51) depende de aspectos que atañen a los datos y al modelo en sí. En la Figura 10-51 el campo *Nivel de medida* permite especificar el nivel de medida de los datos. Las opciones son *Ordinal*, *Intervalo* y *Razón*. Cuando las variables son ordinales, si se selecciona *Desempatar observaciones empata...* se solicitará que sean consideradas como variables continuas, de forma que los empates (valores iguales para casos diferentes) se resuelvan óptimamente. El campo *Condiconalidad* permite especificar qué comparaciones tienen sentido (*Matriz*, *Fila* o *Incondicional*).

El campo *Dimensiones* de la Figura 10-51 permite especificar la dimensionalidad de la solución o soluciones del escalamiento. Se calcula una solución para cada número del rango especificado. Especifique números enteros entre 1 y 6. Se permite un mínimo de 1 sólo si selecciona *Distancia euclídea* como modelo de escalamiento. Para una solución única, especifique el mismo número para el mínimo y el máximo. El campo *Modelo de escalamiento* de la Figura 10-51 permite especificar los supuestos bajo los que se realiza el escalamiento. Las opciones disponibles son *Distancia euclídea* o *Distancia euclídea de diferencias individuales* (también conocida como INDSCAL). Para el modelo de *Distancia euclídea* de diferencias individuales, puede seleccionar *Permitir ponderaciones negativas de sujetos*, si es adecuado para los datos.

El botón *Opciones* de la Figura 10-45 nos lleva a la Figura 10-52, que permite especificar opciones para el análisis de escalamiento multidimensional. El campo *Mostrar* permite seleccionar varios tipos de resultados. Las opciones disponibles son *Gráficos de grupo*, *Gráficos para los sujetos individuales*, *Matriz de datos* y *Resumen del modelo y de las opciones*. El campo *Criterios* permite determinar cuándo debe detenerse la iteración. Para cambiar los valores por defecto, introduzca valores para la *Convergencia de s-stress*, el *Valor mínimo de s-stress* y el *Nº máximo de iteraciones*. El botón *Tratar distancias menores que n como perdidas* lleva a que las distancias menores que *n* se excluyan del análisis.

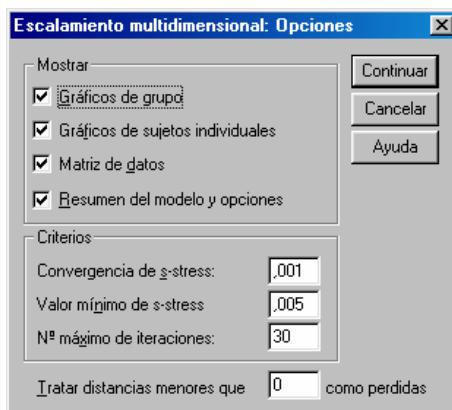


Figura 10-52

Las medidas de disimilitud siguientes son las disponibles para los **datos binarios**:

Distancia euclídea: Se calcula a partir de una tabla 2*2 como $\text{SQRT}(b+c)$, donde b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro.

Distancia euclídea al cuadrado: Se calcula como el número de casos discordantes. Su valor mínimo es 0 y no tiene límite superior.

Diferencia de tamaño: Se trata de un índice de asimetría. Oscila de 0 a 1.

Diferencia de configuración: Medida de disimilitud para datos binarios que oscila de 0 a 1. Se calcula a partir de una tabla 2*2 como $bc/(n^{**2})$, donde b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro y n es el número total de observaciones.

Varianza: Se calcula a partir de una tabla 2x2 como $(b+c)/4n$, donde b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro y n es el número total de observaciones. Oscila de 0 a 1.

Lance y Williams: Se calcula a partir de una tabla 2*2 como $(b+c)/(2a+b+c)$, donde a representa la casilla correspondiente a los casos presentes en ambos elementos y b y c representan las casillas diagonales correspondientes a los casos presentes en un elemento pero ausentes en el otro. Esta medida oscila entre 0 y 1. También se conoce como el coeficiente no métrico de Bray-Curtis.

Si lo desea, puede cambiar los campos *Presente* y *Ausente* para especificar los valores que indican que una característica está presente o ausente. El procedimiento ignorará todos los demás valores.

Las siguientes medidas de disimilaridad están disponibles para **datos de intervalo**:

Distancia euclídea: La raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de los elementos. Ésta es la medida por defecto para datos de intervalo.

Distancia euclídea al cuadrado: La suma de los cuadrados de las diferencias entre los valores de los elementos.

Chebychev: La diferencia absoluta máxima entre los valores de los elementos.

Bloque: La suma de las diferencias absolutas entre los valores de los elementos. También se conoce como la distancia de Manhattan.

Minkowski: La raíz p-ésima de la suma de las diferencias absolutas elevada a la potencia p-ésima entre los valores de los elementos.

Personalizada: La raíz r-ésima de la suma de las diferencias absolutas elevada a la potencia p-ésima entre los valores de los elementos.

Las siguientes medidas de disimilaridad son las disponibles para los **datos de frecuencia**:

Medida de chi-cuadrado: Basado en la prueba de igualdad de chi-cuadrado para dos conjuntos de frecuencias. Ésta es la medida por defecto para datos de recuento.

Medida de Phi-cuadrado: Esta medida es igual a la medida de chi-cuadrado normalizada por la raíz cuadrada de la frecuencia combinada.

Las siguientes opciones están disponibles para la transformación de valores:

Puntuaciones Z: Los valores se estandarizan a una puntuación Z, con una media de 0 y una desviación típica de 1.

Rango -1 a 1: Cada valor del elemento que se tipifica se divide por el rango de los valores.

Rango 0 a 1: El procedimiento sustrae el valor mínimo de cada elemento que se tipifica y después lo divide por el rango.

Magnitud máxima de 1: El procedimiento divide cada valor del elemento que se tipifica por el máximo de los valores.

Media 1: El procedimiento divide cada valor del elemento que se tipifica por la media de los valores.

Desviación típica 1: El procedimiento divide cada valor de la variable o caso que se tipifica por la desviación típica de los valores.

Además, se puede escoger el modo de realizar la tipificación. Las opciones son *Por variable* o *Por caso*.

Una vez elegidas las especificaciones (que se aceptan con el botón *Continuar*), se pulsa el botón *Aceptar* en la Figura 10-45 para obtener los resultados del análisis según se muestra en la Figura 10-53. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 10-54 a 10-60 se presentan varias salidas tabulares y gráficas del procedimiento.



Figura 10-53



Figura 10-54



Figura 10-55



Figura 10-56



Figura 10-57

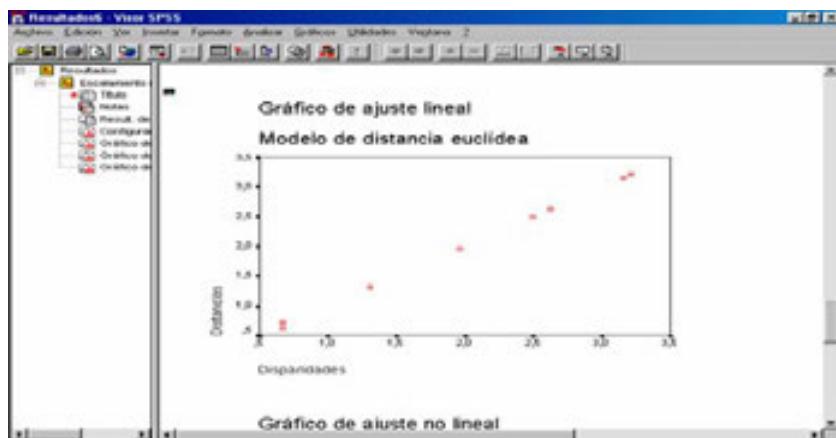


Figura 10-58

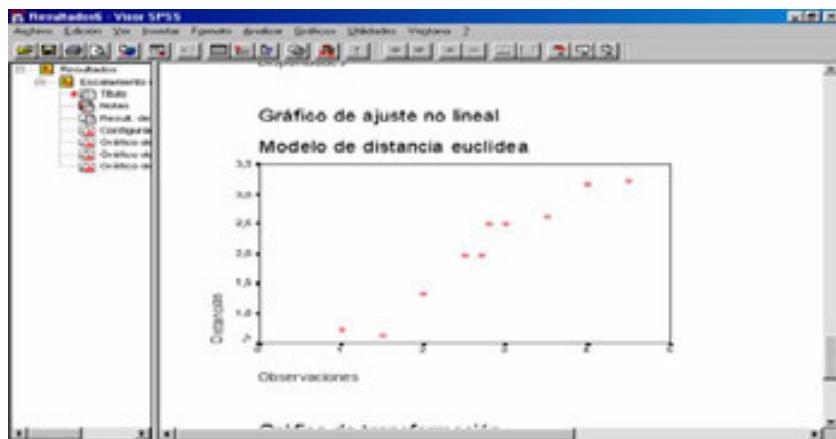


Figura 10-59

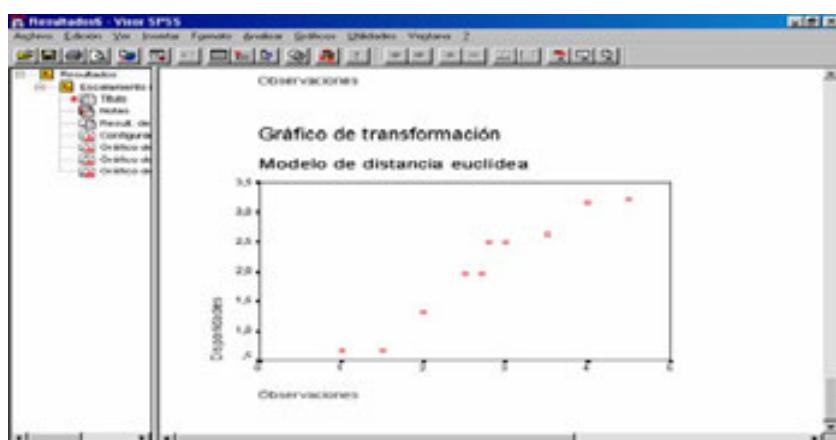


Figura 10-60

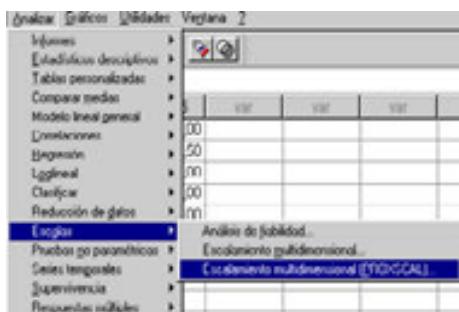
Procedimiento PROXSCAL

El procedimiento PROXSCAL de SPSS, también ejecuta escalamiento multidimensional. El escalamiento multidimensional trata de encontrar la estructura existente en un conjunto de medidas de proximidades entre objetos. Esto se logra asignando las observaciones a posiciones específicas en un espacio conceptual de pocas dimensiones, de modo que las distancias entre los puntos en el espacio concuerden al máximo con las similaridades (o disimilaridades) dadas. El resultado es una representación de mínimos cuadrados de los objetos en dicho espacio de pocas dimensiones que, en muchos casos, le ayudará a entender mejor los datos.

Como ejemplo, el escalamiento multidimensional puede ser muy útil en la determinación de relaciones perceptuales. Por ejemplo, al considerar la imagen de un producto, se puede llevar a cabo un estudio con el fin de obtener un conjunto de datos que describa la similaridad percibida (o proximidad) de este producto con el de la competencia. Mediante estas proximidades y las variables independientes (como el precio), puede intentar determinar las variables que son importantes en la visión que el público tiene del producto y ajustar la imagen de acuerdo con ello.

En cuanto a estadísticos y gráficos, PROXSCAL genera el historial de iteraciones, medidas de stress, descomposición del stress, coordenadas del espacio común, distancias entre objetos dentro de la configuración final, ponderaciones del espacio individual, espacios individuales, proximidades transformadas, variables independientes transformadas, gráficos del stress, diagramas de dispersión del espacio común, diagramas de dispersión de la ponderación del espacio individual, diagramas de dispersión de los espacios individuales, gráficos de transformación, gráficos residuales de Shepard y gráficos de transformación de las variables independientes.

Para obtener un escalamiento multidimensional cargue el fichero de datos (DISTAN1) y elija en los menús (Figura 10-61): *Analizar → Escala → Escalamiento multidimensional (PROXSCAL)*. Accederá al cuadro de diálogo *Formato de datos* (Figura 10-62). Se debe especificar en el campo *Formato de datos* si los datos son medidas de proximidad o si desea crear las proximidades a partir de los datos. En el campo *Número de fuentes*, si los datos son proximidades, debe especificar si dispone de una fuente única o de varias fuentes de medidas de proximidad. Si hay una sola fuente de proximidades, en *Una fuente*, especifique si el conjunto de datos se encuentra en un formato con las proximidades en una matriz a través de las columnas o en una única columna con dos variables diferentes para identificar la fila y la columna de cada proximidad. Si hay varias fuentes de proximidades, en *Varias fuentes*, especifique si el conjunto de datos se encuentra en un formato con las proximidades a través de las columnas en matrices apiladas, en varias columnas con una fuente por cada columna o en una única columna.



El botón *Definir* nos lleva a la Figura 10-63 en la que seleccionaremos al menos tres variables que se utilizarán para crear la matriz de proximidades (o matrices, si hay varias fuentes). Si existen varias variables, seleccione una variable de fuentes (campo *Fuentes*). El número de objetos en la variable de proximidades deberá ser igual al número de filas multiplicado por el número de columnas y por el número de fuentes.



Además, puede definir un modelo para el escalamiento multidimensional, establecer restricciones en el espacio común, establecer criterios de convergencia, especificar la configuración inicial que se va a utilizar y seleccionar gráficos y resultados. El botón *Modelo* (Figura 10-64) permite seleccionar el modelo de escalamiento, la forma de la matriz de proximidades, el tipo de transformaciones que se van a aplicar a las proximidades, el número de dimensiones en la solución y si las proximidades son datos de similaridad o de disimilaridad. Si lo desea, seleccione una medida para crear proximidades. El botón *Restricciones* (Figura 10-65) permite poner restricciones a la solución, fijar algunas coordenadas o hacer que la solución sea una combinación lineal de variables independientes. El botón *Opciones* (Figura 10-66) permite elegir una configuración inicial y seleccionar los criterios de iteración.

El botón *Gráficos* solicita gráficos opcionales, que incluyen stress, espacio común, espacios individuales, ponderaciones de los espacios individuales, proximidades originales respecto a las transformadas, proximidades transformadas respecto a las distancias, variables independientes transformadas y gráficos de las correlaciones entre las dimensiones y las variables (Figura 10-67). El botón *Resultados* permite especificar resultados de tabla pivote y nuevas variables que se van a guardar en archivos externos (Figura 10-68).



Figura 10-65



Figura 10-66



Figura 10-67

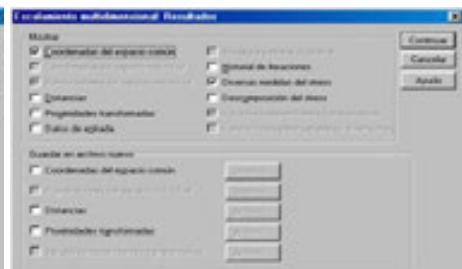


Figura 10-68

Una vez elegidas las especificaciones (que se aceptan con el botón *Continuar*), se pulsa el botón *Aceptar* en la Figura 10-63 para obtener los resultados del análisis EMD según se muestra en la Figura 10-69. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 10-70 y 10-71 se presentan salidas tabulares y gráficas de entre las múltiples que ofrece el procedimiento.



Figura 10-69



Figura 10-70

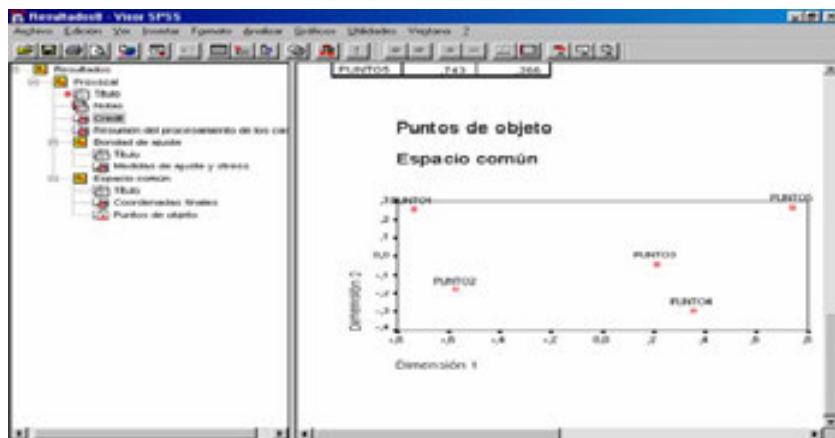


Figura 10-71

Ejercicio 10-1. Consideraremos la matriz de distancias entre 10 ciudades europeas:

Ciudad	Atenas	Berlín	Estocolmo	Londres	Madrid	Moscú	París	Roma	Varsovia	Viena
Atenas	0									
Berlín	1 774	0								
Estoco	2 371	806	0							
Londre	2 355	9 19	1 387	0						
Madrid	2 387	1 855	2 548	1 258	0					
Moscú	2 177	1 565	1 210	2 419	3 371	0				
París	2 065	871	1 516	339	1 048	2 419	0			
Roma	1 048	1 177	1 952	1 419	1 371	2 323	1 097	0		
Varsov	1 581	484	790	1 403	2 258	1 129	1 323	1 290	0	
Viena	1 274	516	1 226	1 210	1 806	1 613	1 016	758	548	0

A partir de estas distancias (archivo 10-1.sav), realizar un escalamiento métrico que sitúe estas ciudades sobre un mapa perceptual que emule el continente europeo.

Comenzamos introduciendo los datos de las distancias entre capitales europeas en el editor de SPSS y a continuación se selecciona *Analizar* → *Escalas* → *Escalamiento multidimensional* (Figura 10-72). Se obtiene la pantalla de entrada del procedimiento de la Figura 10-73. Con los botones *Opciones* y *Modelo* se obtienen pantallas que se llenan como se indica en la Figuras 10-74 y 10-75 (se observa *Razón* en *Nivel de medida*).

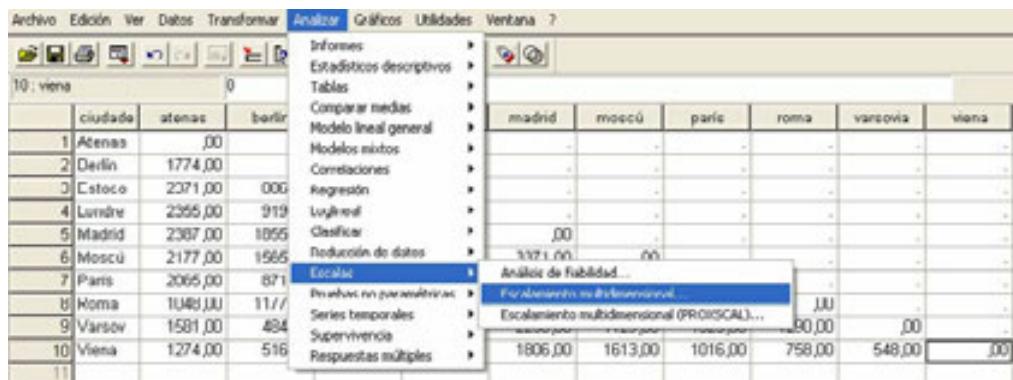


Figura 10-72

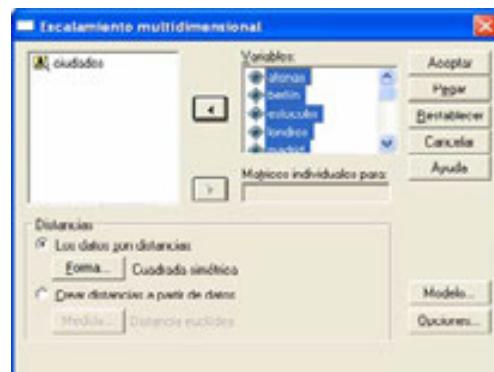


Figura 10-73

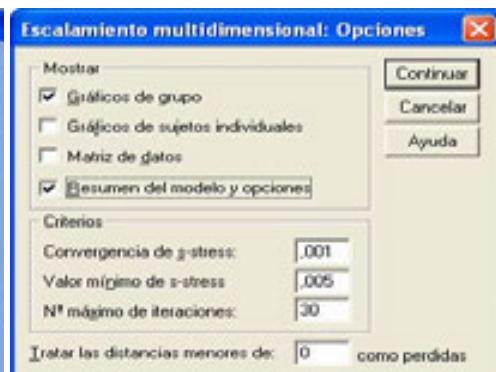


Figura 10-74

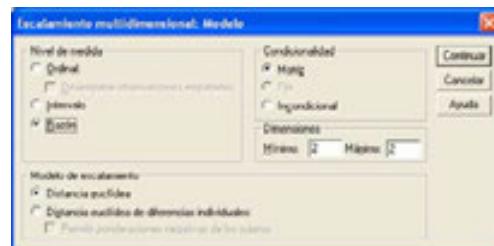


Figura 10-75

Al hacer clic en *Continuar* y *Aceptar*, se obtiene la salida textual del procedimiento ALSCAL que expresa las opciones de datos, de modelo, de salida y de algoritmo, así como el historial de iteraciones y la matriz de coordenadas normalizadas o coordenadas estímulos.

Alscal Procedure Options

Data Options-

Number of Rows (Observations/Matrix)	10
Number of Columns (Variables)	10
Number of Matrices	1
Measurement Level	Ratio
Data Matrix Shape	Symmetric
Type	Dissimilarity
Approach to Ties	Leave Tied
Conditionality	Matrix
Data Cutoff at	,000000

Model Options-

Model	Euclid
Maximum Dimensionality	2
Minimum Dimensionality	2
Negative Weights	Not Permitted

Output Options-

Job Option Header	Printed
Data Matrices	Not Printed
ConFigurations and Transformations	Plotted
Output Dataset	Not Created
Initial Stimulus Coordinates	Computed

Algorithmic Options-

Maximum Iterations	30
Convergence Criterion	,00100
Minimum S-stress	,00500
Missing Data Estimated by	Ulbounds

-

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	,00373	

Iterations stopped because
S-stress is less than ,005000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.

Stress values are Kruskal's stress formula 1.

For matrix
 Stress = ,00352 RSQ = ,99994

— Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension	
		1	2
1	ATENAS	-,1860	1,9206
2	BERLÍN	-,2171	-,3693
3	ESTOCOLM	-,9986	-1,0381
4	LONDRES	,7928	-,9659
5	MADRID	2,1610	-,0867
6	MOSCÚ	-2,2021	-,0454
7	PARÍS	,8881	-,5269
8	ROMA	,6604	,8665
9	VARSOVIA	-,7562	-,0390
10	VIENA	-,1423	,2844

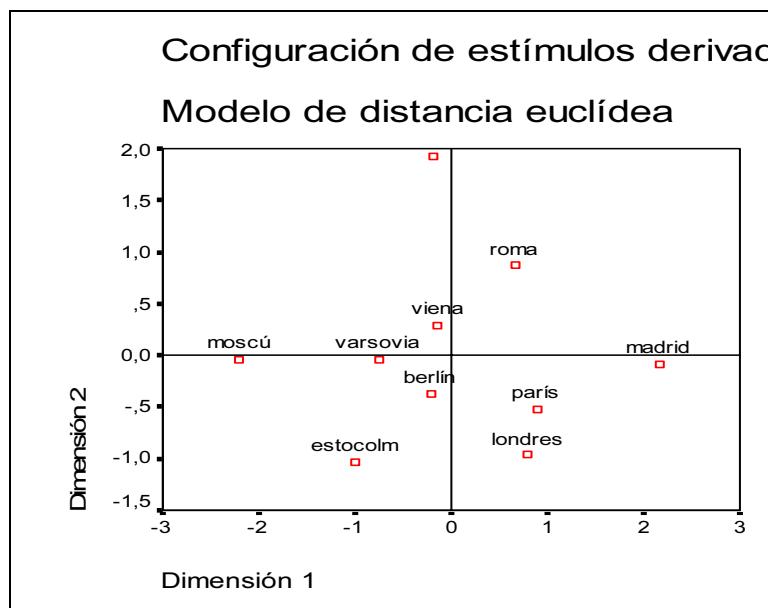


Figura 10-76

Ejercicio 10-2. Se trata de estudiar las relaciones que existen entre 10 tipos diferentes de delitos. Para ello se han formado todos los pares posibles de delitos y se han ordenado estos pares en función de su similitud como sigue:

Delito	Homic	Atraco	Robo	Violaci	Agresi	Desfal	Chant	Secues	Contra	Terr
Homic	0									
Atraco	21	0								
Robo	11	2	0							
Violaci	3	7	9	0						
Agresi	6	4	12	5	0					
Desfalc	45	26	13	40	36	0				
Chantaj	29	28	25	20	22	37	0			
Secues	18	23	16	15	14	41	10	0		
Contrab	34	31	24	30	27	43	42	38	0	
Terroris	8	35	33	32	17	44	19	1	39	0

A partir de esta matriz de similaridades entre delitos (archivo 10-2.sav), realizar un escalamiento no métrico que sitúe estos delitos sobre un mapa perceptual que aclare la clasificación y los relacione convenientemente.

Comenzamos introduciendo los datos de las similitudes entre delitos en el editor de SPSS y a continuación se selecciona *Analizar* → *Escalas* → *Escalamiento multidimensional* (Figura 10-77). Se obtiene la pantalla de entrada del procedimiento de la Figura 10-78. Con los botones *Opciones* y *Modelo* se obtienen pantallas que se llenan como se indica en la Figuras 10-79 y 10-80 (se observa *Ordinal* en *Nivel de medida*).

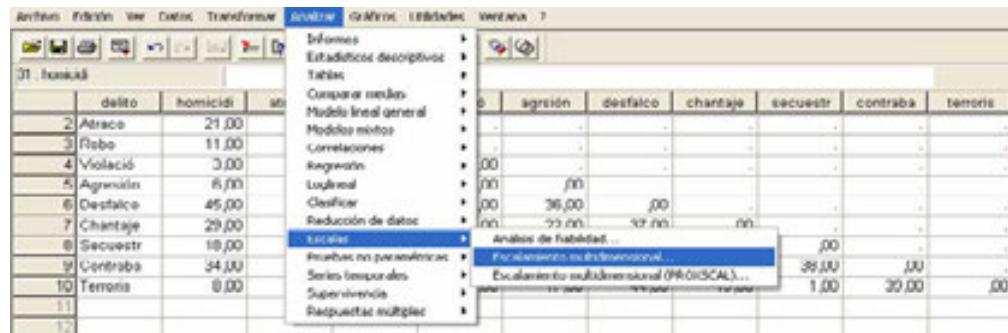


Figura 10-77

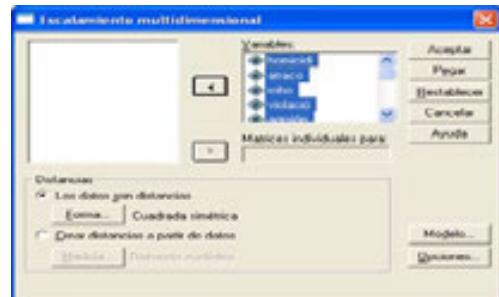


Figura 10-78

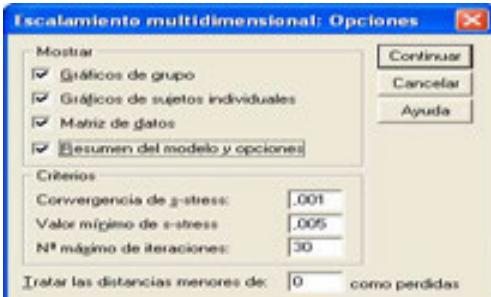


Figura 10-79

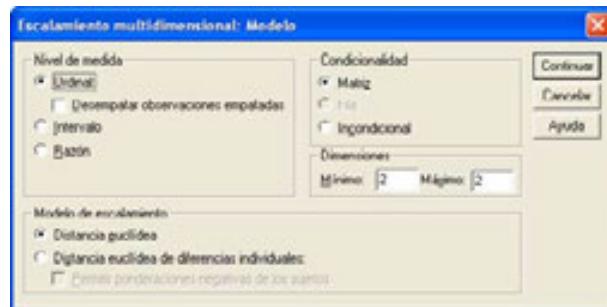


Figura 10-80

El análisis MDS muestra que la solución en dos dimensiones proporciona un buen ajuste (buena convergencia y buenos valores de Stress y RSQ), proporcionando la matriz X de coordenadas en dos dimensiones. La interpretación gráfica de esta matriz se observa en la Figura 10-81

Alscal Procedure Options

Data Options-

Number of Rows (Observations/Matrix)	10
Number of Columns (Variables)	10
Number of Matrices	1
Measurement Level	Ordinal
Data Matrix Shape	Symmetric
Type	Dissimilarity
Approach to Ties	Leave Tied
Conditionality	Matrix
Data Cutoff at	,000000

Model Options-

Model	Euclid
Maximum Dimensionality	2
Minimum Dimensionality	2
Negative Weights	Not Permitted

Output Options-

Job Option Header	Printed
Data Matrices	Not Printed
ConFigurations and Transformations . . .	Plotted
Output Dataset	Not Created
Initial Stimulus Coordinates	Computed

Algorithmic Options-

Maximum Iterations 30
 Convergence Criterion ,00100
 Minimum S-stress ,00500
 Missing Data Estimated by Ulbounds
 Tiestore 45

-

>Number of parameters is 20. Number of data values is 45

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	,15675	
2	,09231	,06444
3	,08345	,00886
4	,08168	,00177
5	,08118	,00050

Iterations stopped because
S-stress improvement is less than ,001000

Stress = ,09688 RSQ = ,95104

-

ConFiguration derived in 2 dimensions

Stimulus Coordinates

Dimension

Stimulus Number	Stimulus Name	1	2
1	HOMICIDI	,9716	,7170
2	ATRACO	-,7309	-,0354
3	ROBO	-,7067	-,1541
4	VIOLACIÓ	,2849	,6861
5	AGRSIÓN	,1973	,5506
6	DESFALCO	-2,2252	-1,3093
7	CHANTAJE	,7028	-1,1680
8	SECUESTR	,9582	-,7533
9	CONTRABA	-1,0016	1,7579
10	TERRORIS	1,5496	-,2915

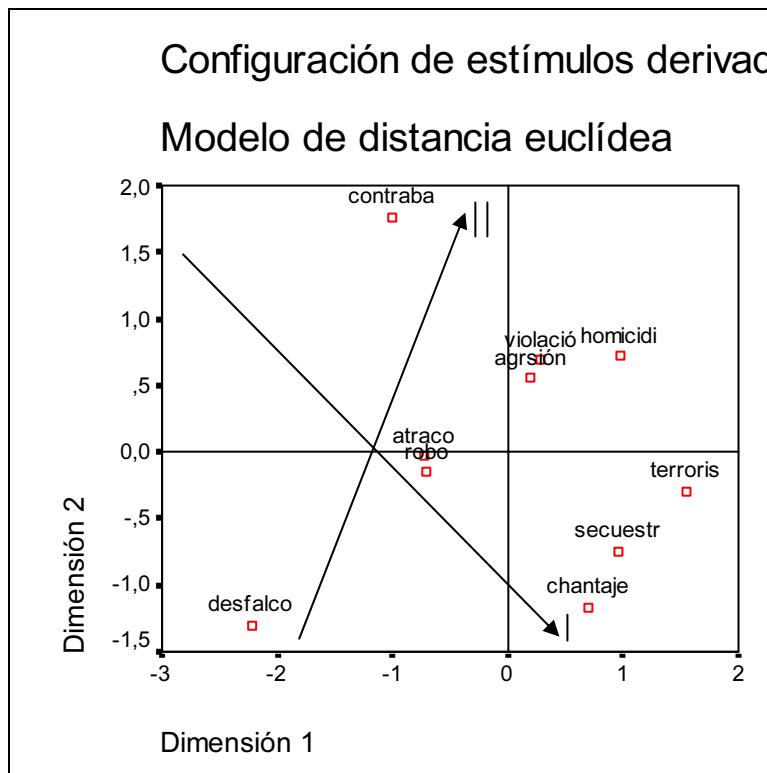


Figura 10-81

Para interpretar el gráfico se han situado dos flechas sobre él. La primera, etiquetada con una barra vertical, muestra una ordenación de los delitos en función de lo personal e impersonal de su naturaleza. Los delitos contra personas (terrorismo, secuestro y chantaje) aparecen en la zona inferior derecha, y a medida que nos desplazamos hacia la zona superior izquierda, encontramos delitos cada vez más impersonales, siendo el más impersonal de todos, el contrabando. La segunda flecha, etiquetada con dos barras verticales, muestra una ordenación de los delitos según su gravedad. Los delitos más graves (homicidio, terrorismo y contrabando) aparecen en la parte más alejada hacia la derecha, y la gravedad desciende a medida que nos desplazamos hacia la izquierda de la gráfica, siendo el delito menos grave el desfalco.

En la Figura 10-82 se presenta el gráfico de transformación de proximidades en rangos originales (de 1 a 45) en disparidades. Cuando el gráfico escalonado es muy brusco (escalones muy diferentes en anchura y separación), el ajuste de disparidades en proximidades es malo, mientras que si los puntos ascienden suavemente hacia la derecha, el ajuste es bueno. En nuestro caso observamos un par de escalones demasiado grandes, lo que indica algún problema en el ajuste, pero tampoco demasiado fuerte como para invalidarlo.

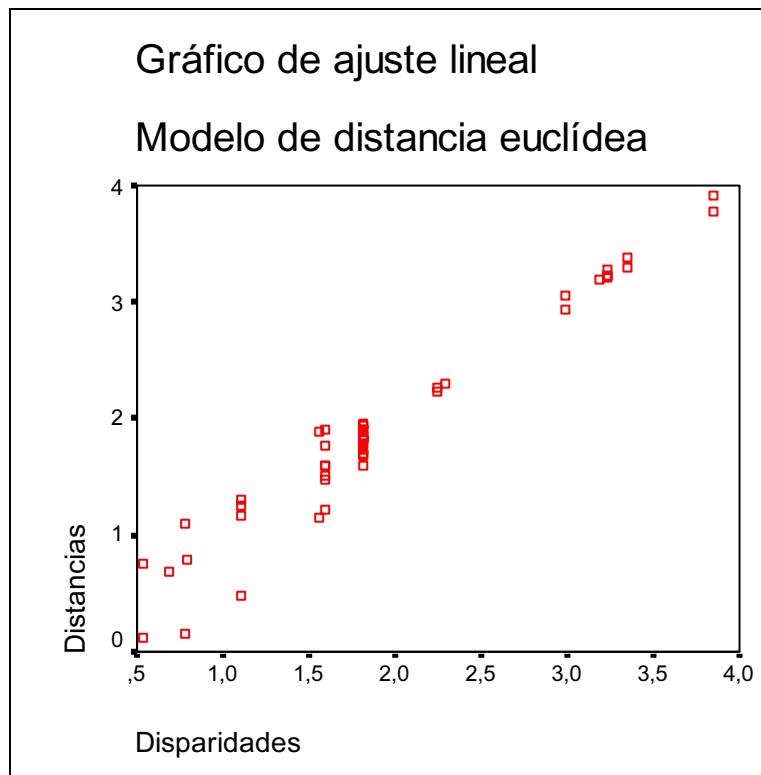


Figura 10-82

Ejercicio 10-3. Se pide a 10 sujetos que ordenen diez tipos de programas de espectáculos diferentes de acuerdo a sus preferencias, resultando los siguientes datos:

Preferencias	Sujeto1	Sujeto2	Sujeto3
<i>Concursos</i>	7	12	12
<i>Documentales</i>	5	3	3
<i>Cine</i>	1	4	7
<i>Humor</i>	6	11	8.
<i>Telediarios</i>	3	1	6
<i>Magazines</i>	8	6	5
<i>Salud</i>	9	5	4
<i>Deportes</i>	12	2	1
<i>Música</i>	2	10	9
<i>Series</i>	11	8	10
<i>Debates</i>	4	7	2
<i>Reality-shows</i>	10	9	11

A partir de estas preferencias (archivo 10-3.sav) derivar las matrices de distancias euclídeas entre los distintos tipos de programas, una para cada sujeto. A partir de estas matrices, realizar un escalamiento no métrico que permita representar estos programas para poder analizarlos, clasificarlos y relacionarlos.

Comenzamos introduciendo la información en el editor de datos de SPSS como 15 columnas (una por cada programa) y 3 filas (una por cada sujeto), es decir situamos sobre el editor de SPSS, la transpuesta de la matriz de preferencias. A continuación utilizamos la opción *Datos* → *Segmentar archivo* (Figura 10-83) y rellenamos la pantalla de entrada como se indica en la Figura 10-84 para segmentar por cada uno de los tres sujetos. Obsérvese que la variable de segmentación es *sujeto*. Se pulsa en *Aceptar* y a continuación se elaboran las tres matrices de distancias mediante la opción *Analizar* → *Correlaciones* → *Distancias* (Figura 10-85) rellenando la pantalla de entrada como se indica en la Figura 10-86. Al hacer clic en *Aceptar* se obtienen las matrices de distancias (Figuras 10-87 a 10-89).

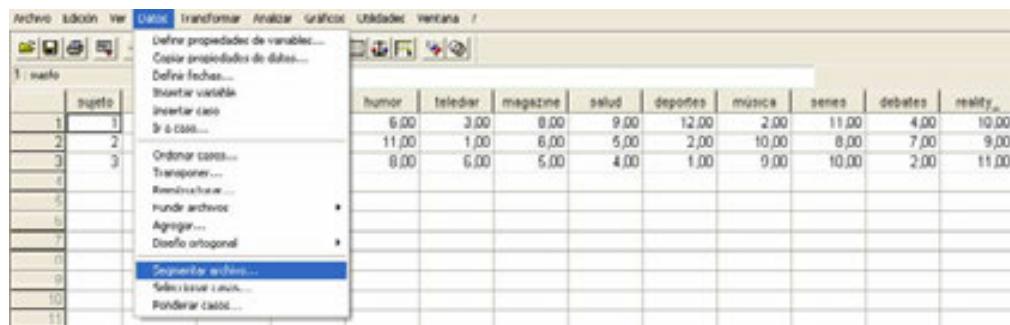


Figura 10-83

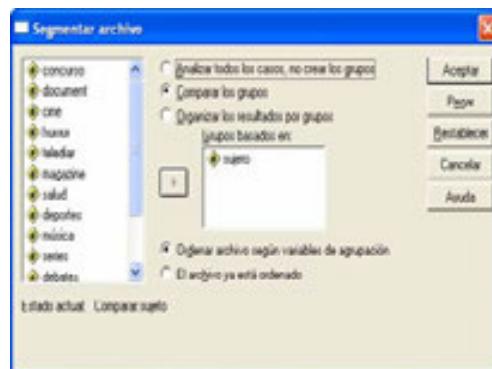


Figura 10-84



Figura 10-85

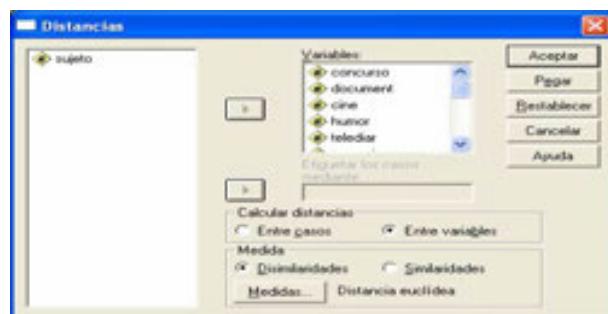


Figura 10-86

Sujeto = Sujeto 1												
	Distancia euclídea											
	CONCURSO	DOCUMENT	CINE	HUMOR	TELECAR	MAGAZINE	SALUD	DEPORTES	MUSICA	SERIES	DEBATES	REALITY
CONCURSO	,000	1,000	6,000	1,000	4,000	1,000	2,000	5,000	5,000	4,000	3,000	3,000
DOCUMENT	2,000	,000	4,000	1,000	2,000	3,000	4,000	7,000	3,000	6,000	1,000	6,000
CINE	6,000	4,000	,000	9,000	2,000	7,000	8,000	11,000	1,000	10,000	3,000	9,000
HUMOR	1,000	1,000	5,000	,000	3,000	2,000	3,000	6,000	4,000	5,000	2,000	4,000
TELECAR	4,000	2,000	2,000	3,000	,000	5,000	6,000	9,000	1,000	8,000	1,000	7,000
MAGAZINE	1,000	3,000	7,000	2,000	5,000	,000	1,000	4,000	6,000	3,000	4,000	2,000
SALUD	2,000	4,000	8,000	3,000	6,000	1,000	,000	3,000	7,000	2,000	5,000	1,000
DEPORTES	5,000	7,000	11,000	6,000	9,000	4,000	2,000	,000	10,000	1,000	8,000	2,000
MÚSICA	5,000	3,000	1,000	6,000	1,000	8,000	7,000	10,000	,000	9,000	2,000	8,000
SERIES	4,000	6,000	10,000	5,000	8,000	3,000	2,000	1,000	9,000	,000	7,000	1,000
DEBATES	3,000	1,000	3,000	2,000	1,000	4,000	5,000	8,000	2,000	7,000	,000	6,000
REALITY	3,000	5,000	9,000	4,000	7,000	2,000	1,000	2,000	9,000	1,000	6,000	,000

Esta es una matriz de dissimilitud.

Figura 10-87

Sujeto = Sujeto 2												
	Distancia euclídea											
	CONCURSO	DOCUMENT	CINE	HUMOR	TELECAR	MAGAZINE	SALUD	DEPORTES	MUSICA	SERIES	DEBATES	REALITY
CONCURSO	,000	8,000	8,000	1,000	11,000	6,000	7,000	10,000	3,000	4,000	5,000	3,000
DOCUMENT	9,000	,000	1,000	8,000	2,000	3,000	2,000	1,000	7,000	5,000	4,000	6,000
CINE	8,000	1,000	,000	7,000	3,000	10,000	5,000	6,000	4,000	3,000	5,000	5,000
HUMOR	1,000	8,000	7,000	,000	12,000	5,000	6,000	9,000	1,000	3,000	4,000	2,000
TELECAR	11,000	2,000	3,000	10,000	,000	5,000	4,000	1,000	9,000	,000	6,000	8,000
MAGAZINE	6,000	3,000	2,000	5,000	5,000	,000	1,000	4,000	4,000	2,000	1,000	3,000
SALUD	7,000	2,000	1,000	6,000	4,000	1,000	,000	3,000	5,000	3,000	2,000	4,000
DEPORTES	10,000	1,000	2,000	9,000	1,000	4,000	3,000	,000	8,000	6,000	5,000	7,000
MÚSICA	2,000	7,000	6,000	1,000	9,000	4,000	5,000	8,000	,000	2,000	3,000	1,000
SERIES	4,000	5,000	4,000	3,000	7,000	2,000	3,000	6,000	2,000	,000	1,000	1,000
DEBATES	5,000	4,000	3,000	4,000	6,000	1,000	2,000	5,000	3,000	1,000	,000	2,000
REALITY	3,000	6,000	5,000	2,000	8,000	3,000	4,000	7,000	1,000	1,000	2,000	,000

Esta es una matriz de dissimilitud.

Figura 10-88

Sujeto = Sujeto 3												
	Distancia euclídea											
	CONCURSO	DOCUMENT	CINE	HUMOR	TELECAR	MAGAZINE	SALUD	DEPORTES	MUSICA	SERIES	DEBATES	REALITY
CONCURSO	,000	8,000	5,000	4,000	8,000	7,000	8,000	11,000	3,000	2,000	10,000	1,000
DOCUMENT	9,000	,000	4,000	5,000	3,000	2,000	1,000	2,000	6,000	2,000	7,000	1,000
CINE	5,000	4,000	,000	1,000	1,000	2,000	3,000	4,000	7,000	1,000	3,000	4,000
HUMOR	4,000	5,000	1,000	,000	2,000	3,000	4,000	7,000	1,000	2,000	8,000	3,000
TELECAR	8,000	3,000	1,000	2,000	,000	1,000	2,000	5,000	3,000	4,000	5,000	9,000
MAGAZINE	7,000	2,000	2,000	3,000	1,000	,000	1,000	4,000	4,000	5,000	3,000	6,000
SALUD	8,000	1,000	3,000	4,000	2,000	1,000	,000	3,000	5,000	6,000	2,000	7,000
DEPORTES	11,000	2,000	6,000	7,000	5,000	4,000	3,000	,000	8,000	9,000	1,000	10,000
MÚSICA	3,000	6,000	2,000	1,000	3,000	4,000	5,000	8,000	,000	1,000	7,000	2,000
SERIES	2,000	7,000	3,000	2,000	4,000	5,000	6,000	9,000	1,000	,000	8,000	1,000
DEBATES	10,000	1,000	5,000	6,000	4,000	2,000	2,000	1,000	7,000	8,000	,000	9,000
REALITY	1,000	8,000	4,000	3,000	5,000	6,000	7,000	10,000	3,000	1,000	8,000	,000

Esta es una matriz de dissimilitud.

Figura 10-89

El siguiente paso será introducir estas tres matrices de distancias euclídeas derivadas de las preferencias de los tres sujetos en el editor de datos de SPSS (una debajo de la otra) para realizar el escalamiento multidimensional no métrico. SPSS permite realizar esta tarea mediante sintaxis de comandos (no mediante menús) ejecutando la sintaxis de la Figura 10-90 (*Ejecutar → Todo*) que se ha escrito en el editor de sintaxis abriendo mediante *Archivo → Nuevo → Sintaxis*.

Ejecutada la sintaxis, se obtiene el fichero *sujetos.sav*, que al cargarlo en el editor de datos presenta el conjunto de las tres matrices de distancias situadas unas debajo de otras (Figura 10-91).

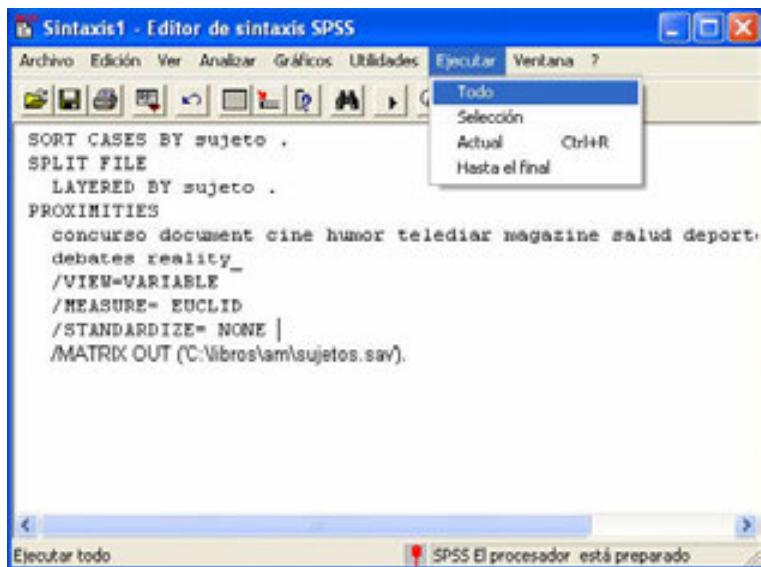


Figura 10-90

sujetos.sav - Editor de datos SPSS														
Archivo Edición Ver Datos Transformar Análisis Gráficos Utilidades Verisión														
T: sujetos														
	idpunto	rowtype	varname	concurs	document	cine	humor	telediar	magazine	salud	deportes	música	séries	del.
1	1	PROX	CONCURS	0,0000	2,0000	6,0000	1,0000	4,0000	1,0000	2,0000	5,0000	5,0000	4,0000	
2	1	PROX	DOCUMEN	2,0000	,0000	4,0000	1,0000	2,0000	3,0000	4,0000	7,0000	3,0000	6,0000	
3	1	PROX	CINE	6,0000	4,0000	,0000	6,0000	2,0000	7,0000	8,0000	11,0000	1,0000	10,0000	
4	1	PROX	HUMOR	1,0000	1,0000	5,0000	,0000	3,0000	2,0000	3,0000	6,0000	4,0000	5,0000	
5	1	PROX	TELEDIAR	4,0000	2,0000	2,0000	3,0000	,0000	5,0000	6,0000	9,0000	1,0000	8,0000	
6	1	PROX	MAGAZINE	1,0000	3,0000	7,0000	2,0000	6,0000	,0000	1,0000	4,0000	6,0000	3,0000	
7	1	PROX	SALUD	2,0000	4,0000	8,0000	3,0000	6,0000	1,0000	,0000	5,0000	7,0000	2,0000	
8	1	PROX	DEPORTE	5,0000	7,0000	11,0000	6,0000	9,0000	4,0000	3,0000	,0000	10,0000	1,0000	
9	1	PROX	MÚSICA	5,0000	3,0000	1,0000	4,0000	1,0000	6,0000	7,0000	10,0000	,0000	9,0000	
10	1	PROX	SERIES	4,0000	6,0000	10,0000	5,0000	8,0000	3,0000	2,0000	1,0000	9,0000	,0000	
11	1	PROX	DEBATES	3,0000	1,0000	3,0000	2,0000	1,0000	4,0000	6,0000	8,0000	2,0000	7,0000	
12	1	PROX	REALITY_	3,0000	5,0000	9,0000	4,0000	7,0000	2,0000	1,0000	2,0000	8,0000	1,0000	
13	2	PROX	CONCURS	0,0000	9,0000	8,0000	1,0000	11,0000	6,0000	7,0000	10,0000	2,0000	4,0000	
14	2	PROX	DOCUMEN	9,0000	,0000	1,0000	8,0000	2,0000	3,0000	2,0000	1,0000	7,0000	5,0000	
15	2	PROX	CINE	6,0000	1,0000	,0000	7,0000	3,0000	2,0000	1,0000	2,0000	6,0000	4,0000	
16	2	PROX	HUMOR	1,0000	8,0000	7,0000	,0000	10,0000	6,0000	6,0000	9,0000	1,0000	3,0000	
17	2	PROX	TELEDIAR	11,0000	2,0000	3,0000	10,0000	,0000	5,0000	4,0000	1,0000	9,0000	7,0000	
18	2	PROX	MAGAZINE	6,0000	3,0000	2,0000	5,0000	5,0000	,0000	1,0000	4,0000	4,0000	2,0000	
19	2	PROX	SALUD	7,0000	2,0000	1,0000	6,0000	4,0000	1,0000	,0000	3,0000	5,0000	3,0000	
20	2	PROX	DEPORTE	10,0000	1,0000	2,0000	9,0000	1,0000	4,0000	3,0000	,0000	8,0000	6,0000	
21	2	PROX	MÚSICA	2,0000	7,0000	6,0000	1,0000	9,0000	4,0000	6,0000	8,0000	,0000	2,0000	
22	2	PROX	SERIES	4,0000	5,0000	4,0000	3,0000	7,0000	2,0000	3,0000	6,0000	2,0000	,0000	
23	2	PROX	DEBATES	5,0000	4,0000	3,0000	4,0000	6,0000	1,0000	2,0000	5,0000	3,0000	1,0000	
24	2	PROX	REALITY_	3,0000	6,0000	5,0000	2,0000	8,0000	3,0000	4,0000	7,0000	1,0000	,0000	
25	3	PROX	CONCURS	0,0000	9,0000	5,0000	4,0000	6,0000	7,0000	6,0000	11,0000	3,0000	2,0000	
26	3	PROX	DOCUMEN	9,0000	,0000	4,0000	5,0000	3,0000	2,0000	1,0000	2,0000	6,0000	7,0000	
27	3	PROX	CINE	6,0000	4,0000	,0000	1,0000	1,0000	2,0000	3,0000	6,0000	2,0000	3,0000	
28	3	PROX	HUMOR	4,0000	5,0000	1,0000	,0000	2,0000	3,0000	4,0000	7,0000	1,0000	2,0000	
29	3	PROX	TELEDIAR	6,0000	3,0000	1,0000	2,0000	,0000	1,0000	2,0000	5,0000	3,0000	4,0000	
30	3	PROX	MAGAZINE	7,0000	2,0000	2,0000	3,0000	1,0000	,0000	1,0000	4,0000	4,0000	5,0000	

Figura 10-91

Para realizar el escalamiento multidimensional no métrico seleccionamos *Analizar* → *Escalas* → *Escalamiento multidimensional* (Figura 10-92). Se obtiene la pantalla de entrada del procedimiento de la Figura 10-93. Con los botones *Opciones* y *Modelo* se obtienen pantallas que se rellenan como se indica en la Figuras 10-94 y 10-95 (se observa *Ordinal* en *Nivel de medida*). Al pulsar *Continuar* y *Aceptar* se obtiene la salida del escalamiento multidimensional. La Figura 10-96 presenta la matriz de coordenadas de los estímulos.

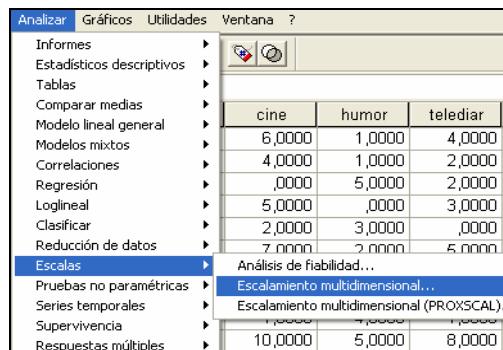


Figura 10-92

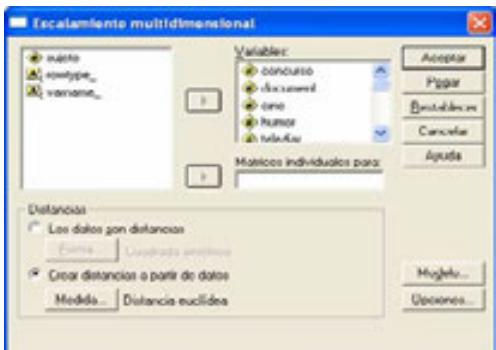


Figura 10-93

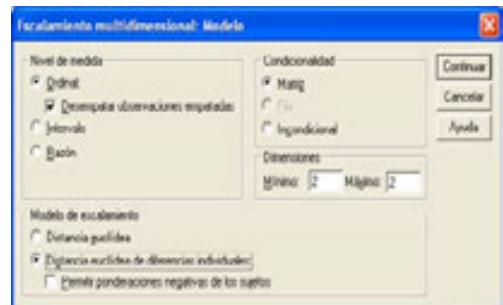


Figura 10-94

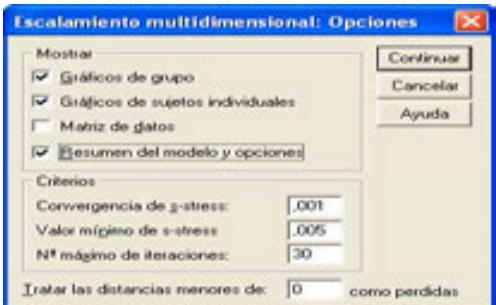


Figura 10-95

Stimulus Coordinates			
Stimulus Number	Stimulus Name	Dimension	
		1	2
1	CONCURSO	1,8378	-,0223
2	DOCUMENT	-1,3293	-,1634
3	CINE	-,5639	-1,1787
4	HUMOR	1,0653	-,2897
5	TELEDIAR	-1,0553	-1,1271
6	MAGAZINE	-,2856	,3597
7	SALUD	-,6119	,8182
8	DEPORTES	-1,5234	1,4272
9	MÚSICA	,9882	-1,1704
10	SERIES	1,0724	,9497
11	DEBATES	-,9513	-,3858
12	REALITY_	1,3568	,7828

Figura 10-96

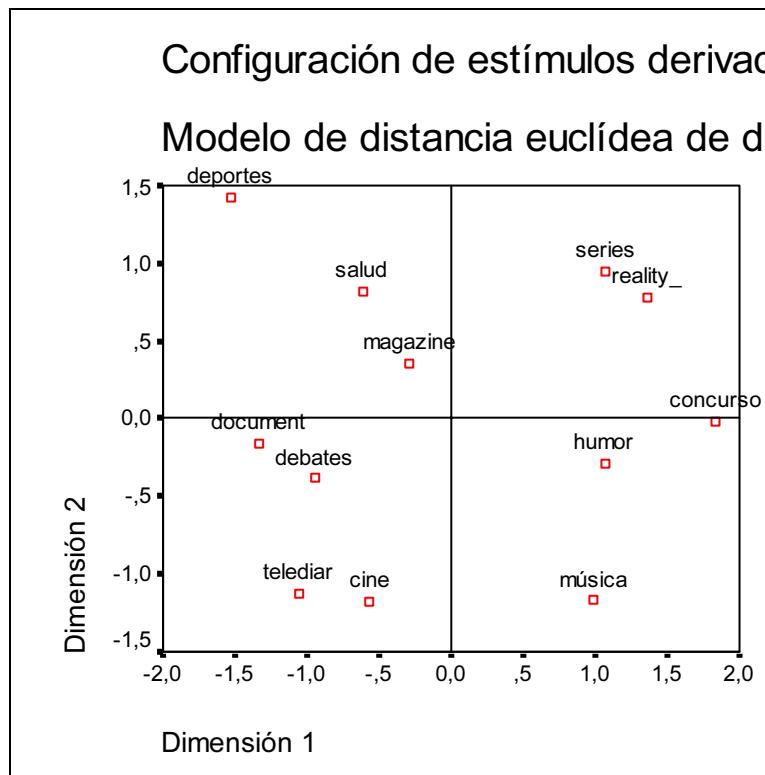


Figura 10-97

La Figura 10-97 representa en el espacio de los estímulos las dos dimensiones de la matriz de coordenadas de los estímulos. Se observa que la primera dimensión distingue a los programas más informativos, situados a la izquierda (deportes, salud, magazines, documentales, debates, telediarios y cine), de los programas no informativos, situados a la derecha (series, reality_show, concursos, programas de humor y música). Por otra parte, la segunda dimensión distingue entre los programas culturales, situados abajo (cine, música, telediarios, debates, documentales) de los programas de entretenimiento, situados arriba (deporte, series, reality_shows, salud, magazines, concursos). Así, mientras los sujetos 2 y 3 clasifican los programas fundamentalmente en función de su grado de información, el sujeto 1 los clasifica prácticamente en función de su contenido cultural.

Ejercicio 10-4. El fichero 10-4.sav contiene los resultados de una encuesta en la que a los individuos encuestados se les pedía manifestar el grado de acuerdo con nueve afirmaciones. Las respuestas se codifican en las nueve variables ítem1 a ítem9 y adicionalmente se clasifican según la variable sexo. Realizar un análisis de no lineal de componentes principales que permita reducir la dimensión de la información original de forma coherente.

Comenzamos cargando en el editor de SPSS los datos del fichero *10-4.sav* mediante *Abrir → datos* y a continuación se selecciona *Anализar → Reducción de datos → Escalamiento óptimo* (Figura 10-98). Se obtiene la pantalla de selección del tipo de escalamiento óptimo que se rellena como se indica en la Figura 10-99 seleccionando CatPCA. Al pulsar en *Definir* se obtiene la pantalla de *Componentes principales categóricas* (Figura 10-100). Con el botón *Resultados* se elige la salida que se desea (Figura 10-101) y con los botones del campo *Gráficos* se elige la salida gráfica (Figura 10-102).

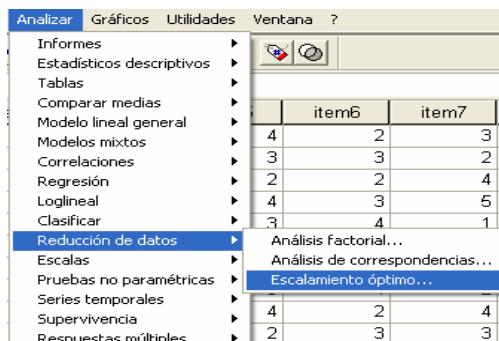


Figura 10-98

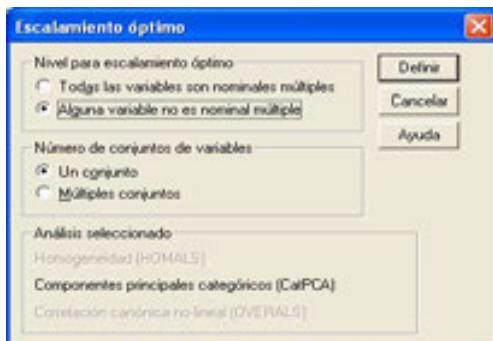


Figura 10-99

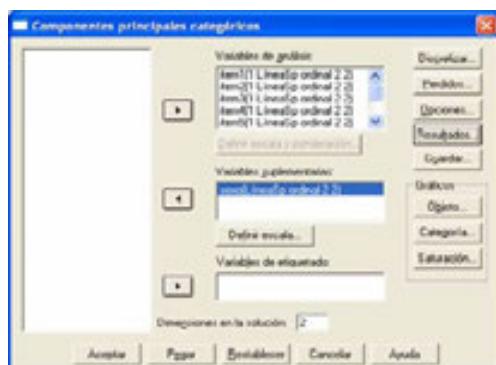


Figura 10-100

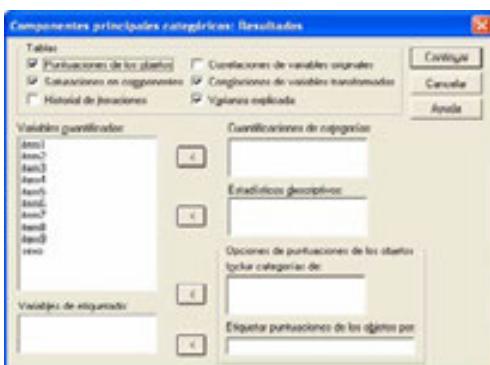


Figura 10-101

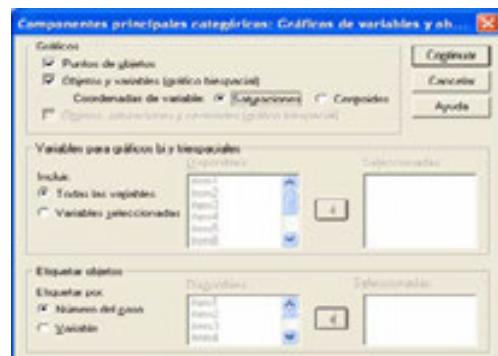


Figura 10-102

Al hacer clic en *Continuar* y *Aceptar* se obtiene la salida del procedimiento de componentes principales categóricas CATPCA (Figura 10-103). En la Figura 10-104 se obtiene la salida resumen del modelo que selecciona dos componentes principales que recogen el 40,538% de la varianza total del modelo. En la Figura 10-105 se ofrece el historial de iteraciones hasta llegar a la solución. En la Figura 10-106 se ve el tanto por ciento de la varianza asociada a cada variable en cada dimensión. En la Figura 10-107 se recogen las cargas o saturaciones de cada una de las variables sobre cada una de las dimensiones del modelo factorial, que representan las proyecciones de cada variable cuantificada en el espacio de los objetos. Se trata del coeficiente de correlación entre cada una de las variables interviniéntes en el modelo con cada una de las dos dimensiones.

Créditos	
CATPCA	
Version 1.1	
by	
Data Theory Scaling System Group (DTSS)	
Faculty of Social and Behavioral Sciences	
Leiden University, The Netherlands	

Figura 10-103

Dimensión	Alfa de Cronbach	Varianza explicada	
		Total (Autovalores)	% de la varianza
1	,557	1,981	22,007
2	,450	1,668	18,531
Total	,817 ^a	3,648	40,538

a. El Alfa de Cronbach Total está basado en los autovalores totales.

Figura 10-104

Historial de iteraciones					
Número de iteración	Varianza explicada		Pérdida		
	Total	Incremento	Total	Coordenadas de centroide	Restricción del centroide a las coordenadas del vector
24 ^a	3,648388	,000008	14,351612	13,736081	,815531

a. Se ha detenido el proceso de iteración debido a que se ha alcanzado el valor de la prueba para la convergencia.

Figura 10-105

	Varianza explicada					
	Coordenadas de centroide		Total (coordenadas del vector)			
	Dimensión		Dimensión		Total	
	1	2	Media	1	2	Total
item1	,065	,475	,270	,014	,456	,469
item2	,286	,293	,289	,266	,235	,501
item3	,056	,588	,322	,003	,573	,575
item4	,272	,075	,173	,254	,022	,276
item5	,381	,076	,228	,357	,007	,364
item6	,400	,124	,262	,346	,112	,458
item7	,398	,092	,245	,375	,067	,441
item8	,270	,031	,153	,264	,015	,279
item9	,111	,205	,188	,103	,181	,284
sexo ^a	,010	,001	,005	,010	,001	,011
Total activo	2,299	1,965	2,132	1,981	1,688	3,648
% de la varianza	25,542	21,834	23,688	22,007	18,531	40,538

a. Variable suplementaria.

Figura 10-106

	Saturaciones en componentes		
	Dimensión	1	2
item1		-,117	,675
item2		,515	-,485
item3		,055	,757
item4		-,504	-,149
item5		-,597	-,086
item6		,588	-,335
item7		,612	,258
item8		,514	,124
item9		,321	,425
sexo ^a		,100	,025

Normalización principal por variable.

a. Variable suplementaria.

Figura 10-107

En cuanto a las salidas gráficas del procedimiento, en la Figura 10-108 se presenta el gráfico de saturaciones en las componentes que se utiliza para agrupar nuestras variables en las dos componentes. Está claro que *item2* e *item6* se asocian con una primera componente e ítem4 e ítem5 con la segunda componente. Pero ya no está tan claro con qué componente principal asociar el resto de las variables. Según la Figura podría ser lógico asociarlas todas con la primera componente. También podrían asociarse *item7* e *item8* con la primera componente e *item1*, *item3* e *item9* con la segunda.

No obstante esta clasificación de las variables en componentes también puede realizarse observando la tabla de saturaciones en las componentes de la Figura 10-107. Se observa en esta tabla que para la componente 2, las saturaciones más altas las presentan las variables *ítem1*, *ítem3* e *ítem9*. Para la componente 1 las saturaciones más altas las presentan *ítem2*, *ítem6*, *ítem4*, *ítem5* e *ítem8* (*ítem4* e *ítem5* con valor negativo, por eso aparecen a la izquierda del gráfico). Luego la forma definitiva de agrupar las variables en componentes sería asociar las variables *ítem4*, *ítem5*, *ítem2*, *ítem6*, *ítem7* e *ítem8* en una componente y las variables *ítem1*, *ítem3* e *ítem9* en la otra componente, siendo las asociaciones más indefinidas las de las variables *ítem7* e *ítem8*. Se observa que la mejor forma de asociar las variables a las componentes principales es analizar simultáneamente la tabla de las saturaciones en las componentes de la Figura 10-107 y el gráfico de las saturaciones en las componentes de la Figura 10-108. La Figura 10-109 presenta la gráfica de puntuaciones de los objetos etiquetadas por el número de caso y en la Figura 10-110 se observa el gráfico de dispersión biespacial, que muestra sobre el mismo gráfico las puntuaciones de los objetos etiquetadas por el número de caso y las saturaciones en las componentes.

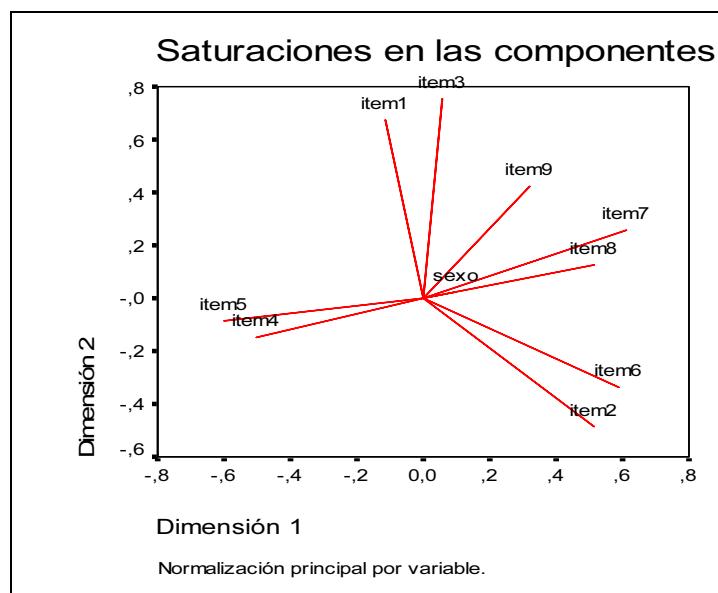


Figura 10-108

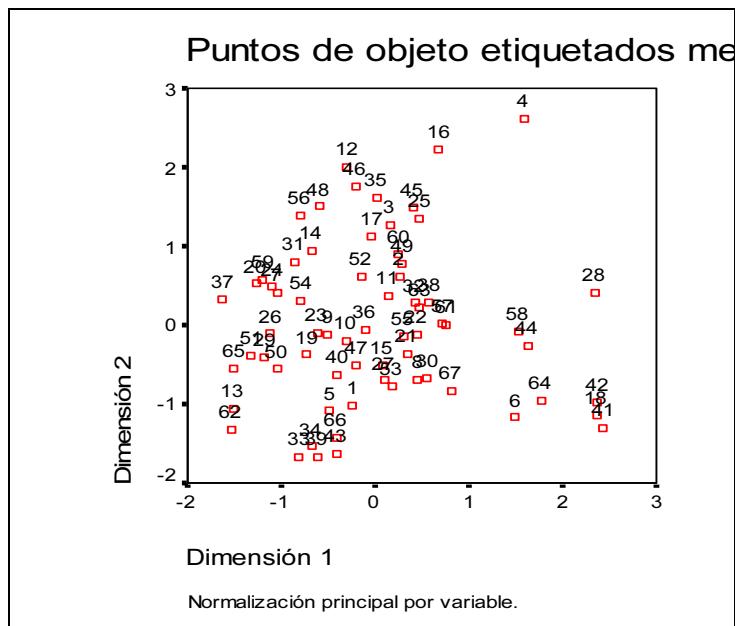


Figura 10-109

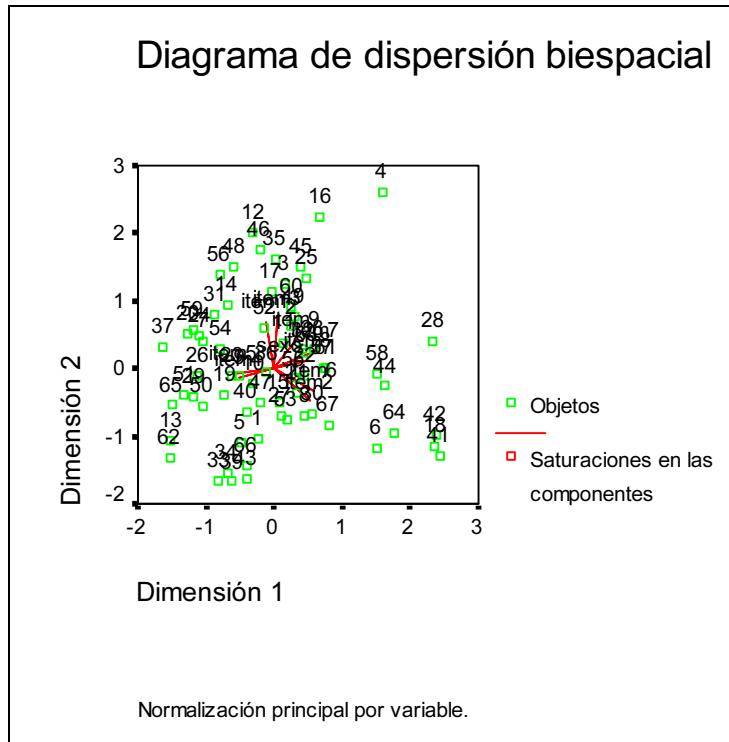


Figura 10-110

Ejercicio 10-5. El fichero 10-5.sav contiene los resultados de una encuesta en la que a los individuos encuestados se les pedía manifestar el grado de acuerdo con nueve afirmaciones. Las respuestas se codifican en las nueve variables ítem1 a ítem9 y adicionalmente se clasifican según la variable sexo. Realizar un análisis no lineal de correlación canónica tomando como primer conjunto de variables ítem1, ítem4 e ítem6, y como segundo conjunto de variables ítem2, ítem3 e ítem5.

Comenzamos cargando en el editor de SPSS los datos del fichero 10-5.sav mediante Abrir → Datos y a continuación se selecciona Analizar → Reducción de datos → Escalamiento óptimo. Se obtiene la pantalla de selección del tipo de escalamiento óptimo que se rellena como se indica en la Figura 10-111 seleccionando OVERALS (*Múltiples conjuntos*). Al pulsar en Definir se obtiene la pantalla de Análisis de correlación canónica no lineal (Figura 10-112) en cuyo campo Variables se introducen el primer conjunto de variables para el análisis. Con el botón Definir rango y escala se declara el máximo y el mínimo de la escala de medida (Figura 10-113). Se hace clic en Continuar y ya se tiene definido el primer conjunto de variables (Figura 10-114). Se hace clic en Siguiente y se introduce el segundo conjunto de variables definiendo también su rango y escala (Figura 10-115). Con el botón Opciones se elige la salida que se desea para el análisis, tanto tabular como gráfica (Figura 10-116). Se hace clic en Continuar y en Aceptar, con lo que ya tenemos la salida del procedimiento OVERALS.

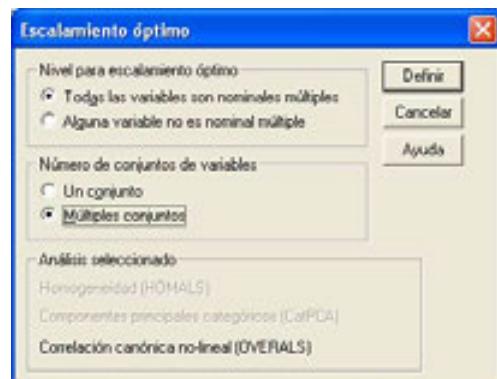


Figura 10-111

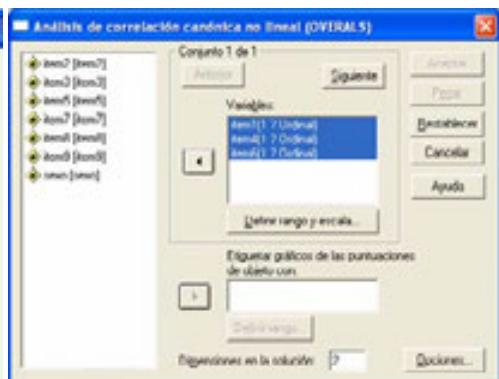


Figura 10-112

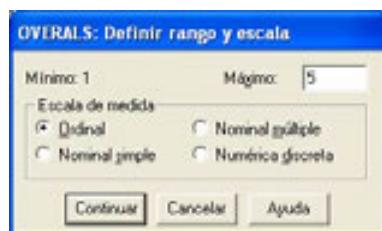


Figura 10-113

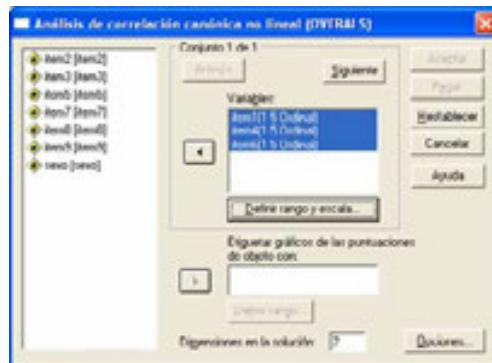


Figura 10-114



Figura 10-115

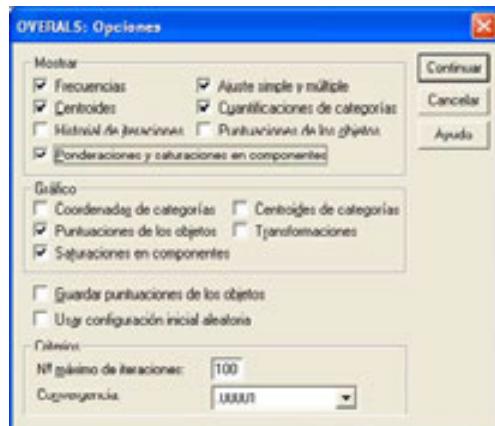


Figura 1-116

La salida tabular comienza ofreciendo listado de las variables con los dos grupos de variables que intervienen en el análisis junto a su número de categorías (Figura 10-117), el historial de iteraciones y el resumen del análisis (Figura 10-118). El historial de iteraciones presenta un informe sobre las iteraciones sucesivas que se llevan a cabo para establecer una relación canónica entre los dos conjuntos. Este proceso de búsqueda de una solución que satisfaga el valor de convergencia (llamado valor del test de la convergencia) desemboca en el cálculo de un valor de pérdida y otro de ajuste para la iteración 0 y la iteración en la que se produce la convergencia (la 77 en nuestro caso). También se presenta la diferencia entre las dos últimas iteraciones (0,000007). En el resumen del análisis, OVERALS muestra la pérdida por cada conjunto en cada dimensión. La suma de las pérdidas del conjunto 1 y del conjunto 2 deben de coincidir. La pérdida media por dimensiones indica una pérdida moderada (0,580). El ajuste de la prueba representa un valor alto (1,420) y los autovalores (0,751 y 0,670) muestran una distribución de cargas de explicación de la varianza del modelo algo superior en la dimensión 1 que en la 2. La Figura 10-119 presenta la tabla de ponderaciones y la de saturaciones en las componentes.

Creditos		
OVERALS		
Version 1.0		
by		
Data Theory Scaling System Group (DTSS)		
Faculty of Social and Behavioral Sciences		
Leiden University, The Netherlands		
Resumen del procesamiento de los casos		
Casos usados en el análisis	67	
Lista de variables		
Conjunto	Número de categorías	Nivel de escalamiento óptimo
1	item1 item4 item6	5 Ordinal 5 Ordinal 5 Ordinal
2	item2 item3 item5	5 Ordinal 5 Ordinal 5 Ordinal

Figura 10-117

Historial de iteraciones			
	Pérdida	Ajuste	Diferencia desde la iteración anterior
0a	,798666	1,201334	
77 b	,579106	1,420894	,000007

a. La pérdida de la iteración 0 es la pérdida de la solución con todas las variables simples tratadas como numéricas (con una diferencia de pérdida de 0,0001 y un número máximo de 50 iteraciones).

b. Se ha detenido el proceso de iteración debido a que se ha alcanzado el valor de la prueba para la convergencia.

Resumen del análisis

	Dimensión			Suma
		1	2	
Pérdida	Conjunto 1	,250	,331	,581
	Conjunto 2	,249	,329	,579
	Media	,249	,330	,580
Autovalores		,751	,870	
Ajuste				1,420

Figura 10-118

Ponderaciones		
Conjunto	Dimensión	
	1	2
1 item1	,514	,671
item4	-,370	,380
item6	,584	-,351
2 item2	,182	-,838
item3	-,206	-,384
item5	-,777	-,359

Saturaciones en componentes		
Conjunto	Dimensión	
	1	2
1 item1a,b	,530	,638
item4a,b	-,436	,273
item6a,b	,542	-,389
2 item2a,b	,436	-,665
item3a,b	-,190	-,201
item5a,b	-,814	-,102

a. Nivel de escalamiento óptimo: Ordinal
b. Proyecciones de las variables cuantificadas simples en el espacio de los objetos

Figura 10-119

La tabla de ponderaciones muestra los pesos por cada dimensión desglosados por un grupo de ítems del primer análisis y por sus respectivos elementos. Se puede observar la elevada fuerza explicativa del ítem 5 dentro de la dimensión 1 y de la carga del ítem 2 en la dimensión 2. Estas ponderaciones o pesos representan los coeficientes de correlación de cada dimensión para todas las variables cuantificadas de un conjunto, donde las puntuaciones de los objetos efectúan un análisis de la regresión sobre las variables cuantificadas.

La tabla de saturaciones en las componentes contempla las cargas de las componentes por variables simples, es decir las proyecciones de las variables cuantificadas en el espacio de los objetos. Estas cargas son una indicación de la contribución de cada variable a la dimensión dentro de cada conjunto. Se aprecia la elevada fuerza explicativa del ítem 3, así como las de los ítems 1 y 2. El gráfico de saturaciones en componentes (Figura 10-122) representa en el plano de las dos dimensiones las cargas de las componentes para variables simples. Como ya hemos dicho, se observa la elevada fuerza explicativa del ítem 3 en la dimensión 1, así como la de los ítems 1 y 2 en la dimensión 2. La tabla de ajuste de la Figura 10-121 resume datos de ajuste múltiple, simple y perdida simple por dimensiones para cada variable de cada uno de los conjuntos del análisis.

		Ajuste					
Conjunto	ítem	Ajuste múltiple		Ajuste simple		Pérdida simple	
		Dimensión		Dimensión		Dimensión	
		1	2	1	2	1	2
1	item1 ^a	,283	,456	,739	,265	,451	,715
	item4 ^a	,163	,147	,310	,137	,144	,281
	item5 ^a	,355	,140	,495	,341	,123	,464
2	item2 ^a	,064	,705	,770	,033	,703	,736
	item3 ^a	,053	,160	,213	,043	,147	,190
	item6 ^a	,605	,137	,742	,604	,129	,733

a. Nivel de escalamiento óptimo: Ordinal

Figura 10-121

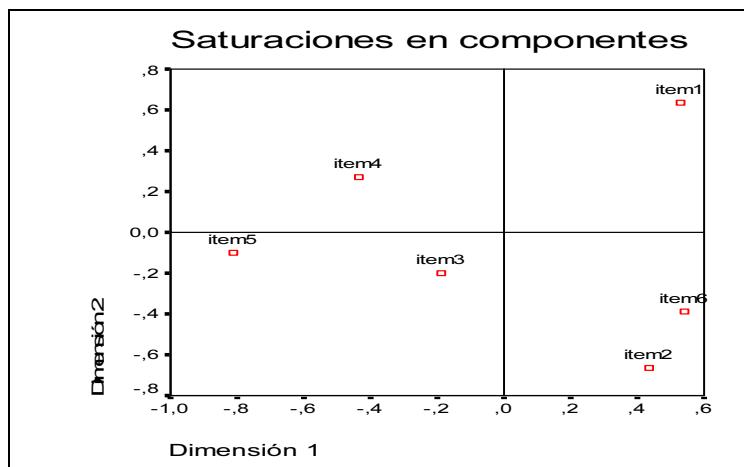


Figura 10-122

MODELOS LOGARÍTMICO LINEALES Y TABLAS DE CONTINGENCIA

TABLAS DE CONTINGENCIA

Consideramos una población (o una muestra) compuesta por N individuos sobre los que se pretende analizar simultáneamente dos atributos o factores (variables cualitativas). Designemos por A_1, \dots, A_h y por B_1, \dots, B_k las h y k modalidades del factor A y del factor B respectivamente, y por n_{ij} el número de individuos que presentan a la vez las modalidades A_i y B_j . La tabla estadística que describe estos N individuos, denominada **tabla de contingencia**, será una tabla de doble entrada como la siguiente:

$A, B \rightarrow$ \downarrow	B_1	B_2	\cdots	B_j	\cdots	B_k	$n_{i\cdot}$
A_1	n_{11}	n_{12}	\cdots	n_{1j}	\cdots	n_{1k}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	\cdots	n_{2j}	\cdots	n_{2k}	$n_{2\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_i	n_{i1}	n_{i2}	\cdots	n_{ij}	\cdots	n_{ik}	$n_{i\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_h	n_{h1}	n_{h2}	\cdots	n_{hj}	\cdots	n_{hk}	$n_{h\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	\cdots	$n_{\cdot j}$	\cdots	$n_{\cdot k}$	$n_{\cdot \cdot}$

Distribuciones marginales y condicionadas

Al igual que en el caso de las variables cuantitativas, en esta tabla $n_{i\cdot}$ y $n_{\cdot j}$ nos proporcionan las **frecuencias marginales**, es decir, el número de veces que aparece la modalidad i -ésima de A , con independencia de cuál sea la modalidad de B , es $n_{i\cdot}$, y el número de veces que aparece la modalidad j -ésima de B , independientemente de cuál sea la modalidad de A con el que se da conjuntamente B , es $n_{\cdot j}$. De esta forma tenemos que las **distribuciones marginales** de A y B vienen dadas por $(A_i; n_{i\cdot})$ y $(B_j; n_{\cdot j})$.

A partir de la tabla de contingencia es posible formar un nuevo tipo de distribuciones, que denominaremos **distribuciones condicionadas** debido a que para su obtención es preciso definir previamente una condición. Esta condición hará referencia a la fijación a priori de una modalidad (o modalidades) de una de las variables cualitativas o factores, para posteriormente calcular la distribución de la otra variable cualitativa sujeta a esa condición. Si fijamos la variable B en el valor B_2 (podríamos fijar más de un único valor), la distribución de la variable A condicionada a que B tome el valor B_2 vendrá dada por:

$$\begin{array}{c} A / B_2 \quad n_i / 2 \\ \hline A_1 & n_{12} \\ A_2 & n_{22} \\ \vdots & \vdots \\ A_h & n_{h2} \\ \hline & n_{\cdot 2} \end{array}$$

Donde A/B_2 nos dará los valores que puede tomar la variable A cuando la B toma el valor B_2 , y $n_i/2$ nos da las frecuencias con que se presentan cada uno de los valores (modalidades).

En general, dado que se pueden establecer condiciones sobre A y B calculando posteriormente la distribución de A o B sujeta a esa condición, nos encontramos distribuciones que, de manera genérica, tendrán la forma:

$$\begin{array}{cc} A / B_j \quad n_i / j & B / A_i \quad n_j / i \\ \hline A_1 & n_{1j} & B_1 & n_{i1} \\ A_2 & n_{2j} & B_2 & n_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ A_h & n_{hj} & B_k & n_{ik} \\ \hline & n_{\cdot j} & & n_{i\cdot} \end{array}$$

Para todas las distribuciones será posible trabajar con frecuencias relativas en vez de con frecuencias absolutas.

INDEPENDENCIA Y ASOCIACIÓN DE VARIABLES CUALITATIVAS. COEFICIENTES

En el caso de variables cualitativas la falta de independencia suele denominarse asociación, y el análisis del grado de asociación entre variables cualitativas tiene fuerte incidencia en la estadística de atributos.

Ya hemos visto que de forma análoga al caso de dos variables cuantitativas, la observación simultánea de dos atributos da lugar a una tabla de doble entrada, en donde n_{ij} indica el número de objetos o individuos que poseen conjuntamente las modalidades indicadas en la fila i -ésima y en la columna j -ésima de la tabla de contingencia. También hemos visto que las distribuciones que se refieren a uno solo de los dos atributos o variables cualitativas también se denominan distribuciones marginales.

Se dice que dos atributos A y B son independientes cuando entre ellos no existe ningún tipo de influencia mutua. Si dos atributos, A y B , son independientes estadísticamente, la frecuencia relativa conjunta será igual al producto de las frecuencias marginales respectivas. Para que A y B sean independientes habrá de cumplirse que $n_{ij} = (n_i \cdot n_j)/N$ para todo i, j . En la práctica basta con que la relación se verifique para $(h-1)(k-1)$ valores de n_{ij} , ya que entonces se verificará para todos los restantes.

Si designamos por n_{ij} la frecuencia conjunta correspondiente a las modalidades A_i del atributo A y B_j de B , y por n_{ij}' la frecuencia teórica que correspondería en el caso de que ambos atributos fuesen independientes, esto es, $n_{ij}' = (n_i \cdot n_j)/N$, $i = 1, \dots, h$, $j = 1, \dots, k$, siendo N el total de elementos que se estudian, definimos el **coeficiente de contingencia χ^2** como sigue:

$$\chi^2 = \sum_{i=1}^h \sum_{j=1}^k \frac{(n_{ij}' - n_{ij})^2}{n_{ij}'}$$

Este coeficiente también se denomina en la literatura estadística **cuadrado de la contingencia**, y puede expresarse de forma más sencilla para el cálculo como sigue:

$$\chi^2 = \sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}^2}{n_{ij}} - N$$

El coeficiente de contingencia χ^2 se utiliza para realizar un contraste formal para la hipótesis nula de independencia de los atributos A y B cuya información muestral se recoge en la tabla de contingencia dada. La hipótesis alternativa es la existencia de asociación entre los atributos A y B . El contraste se basa en que, bajo la hipótesis nula de independencia de los atributos A y B , el estadístico χ^2 se distribuye según una Chi cuadrado con $(h-1)(k-1)$ grados de libertad.

Para realizar el contraste se halla el valor k tal que $P(\chi^2_{(h-1)(k-1)} \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico χ^2 para los datos dados de la tabla de contingencia es mayor que k se rechaza la hipótesis nula de independencia de los atributos A y B al nivel fijado α . En caso contrario se acepta la independencia.

Cuando el tamaño muestral es pequeño (N menor que 150) se utiliza el test exacto de Fisher para contrastar la independencia de atributos. En este caso suele introducirse una corrección por continuidad en el estadístico de la Chi-cuadrado, tomando en su lugar para el contraste de independencia el estadístico corregido de Yates, cuya expresión es la siguiente:

$$\chi^2 = \sum_{i=1}^h \sum_{j=1}^k \frac{(|n'_{ij} - n_{ij}| - \frac{1}{2})^2}{n'_{ij}}$$

Como concepto contrario al de independencia tenemos el de **asociación**. Se dice que A y B están asociados cuando aparecen juntos en mayor número de casos que el que cabría esperar si fuesen independientes. Según que esa tendencia a coincidir o no coincidir esté más o menos marcada, tendremos distintos grados de asociación. Para medirlos se han ideado diversos procedimientos, denominados **coeficientes de asociación**, entre los que destacaremos los siguientes:

- **Cuadrado medio de la contingencia:** Se trata de una medida de asociación sencilla, que no es más que el cociente entre el coeficiente de contingencia χ^2 y el tamaño de la muestra N , con lo cual, se elimina el efecto del tamaño muestral. Este coeficiente alcanza el valor máximo 1 cuando entre los dos atributos existe asociación perfecta estricta. El valor del coeficiente es cero si los atributos son independientes. Se trata de una medida muy sensible a la presencia de totales marginales desequilibrados, por lo cual, cuando esta circunstancia se presenta, los valores tomados por esta medida pueden llevarnos a conclusiones falsas. Tanto el coeficiente de contingencia como el cuadrado medio de la contingencia no pueden ser nunca negativos. La expresión del cuadrado medio de la contingencia será :

$$\Phi^2 = \chi^2/N = \frac{1}{N} \sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}^2}{n'_{ij}} - 1$$

- **Coeficiente de contingencia C de K. Pearson:** Se trata de un coeficiente definido como $C = (\chi^2/(N+\chi^2))^{1/2}$. El coeficiente C tiene un campo de variación entre 0 y 1, de manera que su valor es cero cuando existe una carencia absoluta de asociación entre los atributos, o sea, cuando los atributos son independientes. Cuando los atributos muestran una total asociación entre sí, el coeficiente se aproxima a 1, pero sólo se alcanzaría el valor 1 en el caso ideal de infinitas modalidades. Se puede demostrar que en el caso de una tabla de contingencia cuadrada ($h=k$), el límite superior de C es $S = ((h-1)/h)^{1/2}$, lo que permitiría calcular un nuevo valor para esta medida, llamado coeficiente ajustado, que vendría dado por $C_A = C/S$. Este coeficiente ajustado podría resultar de interés, puesto que proporciona una idea del verdadero grado de asociación, al evaluar la discrepancia entre el valor obtenido y el máximo que podría alcanzar para la tabla dada. La expresión del coeficiente de contingencia C de K. Pearson será:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{\Phi^2}{\Phi^2 + 1}} = \sqrt{1 - \frac{N}{\sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}^2}{n_{ij}}}}$$

- **El coeficiente T de Tschuprow:** Se trata de un coeficiente que depende de χ^2 , del número de filas y columnas de la tabla de contingencia y del total de elementos N . El coeficiente varía entre 0 y 1, pero no alcanza el máximo valor cuando la tabla analizada es rectangular, aunque sí cuando la tabla es cuadrada. La expresión de este coeficiente es la siguiente:

$$T = \sqrt{\frac{\chi^2 / N}{\sqrt{(h-1)(k-1)}}} = \sqrt{\frac{\Phi^2}{\sqrt{(h-1)(k-1)}}} = \sqrt{\frac{C^2}{(1-C^2)\sqrt{(h-1)(k-1)}}}$$

También se cumple que:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{\Phi^2}{\Phi^2 + 1}} = \sqrt{\frac{T^2 \sqrt{(h-1)(k-1)}}{1 + T^2 \sqrt{(h-1)(k-1)}}}$$

- **El coeficiente V de Cramer:** Se trata de un coeficiente que toma el valor 1 cuando existe asociación perfecta entre atributos, cualquiera que sea el número de filas y columnas de la tabla de contingencia analizada. Cuando la tabla es cuadrada $V = T$, y en caso contrario, $V > T$. Su expresión es:

$$V = \sqrt{\frac{\Phi^2}{m}} = \sqrt{\frac{\chi^2}{mN}} \quad \text{donde } m = \min(h-1, k-1)$$

- **Coeficientes Lambda de Goodman y Kruskall:** Se trata de coeficientes que ya no dependen de χ^2 . Suponiendo que se ha elegido Y como factor explicado y X como explicativo, se evalúa la capacidad de X para predecir Y mediante el coeficiente λ_y , cuya expresión es:

$$\lambda_y = \frac{\sum_{i=1}^h \max_j n_{ij} - \max_j n_{.j}}{N - \max_j n_{ij}}$$

De la misma forma, suponiendo que se ha elegido X como factor explicado e Y como explicativo, se evalúa la capacidad de Y para predecir X mediante el coeficiente λ_x , cuya expresión es:

$$\lambda_x = \frac{\sum_{j=1}^k \max_i n_{ij} - \max_i n_{i.}}{N - \max_i n_{i.}}$$

Tanto λ_x como λ_y varían entre 0 y 1 y están especialmente pensadas como medidas asimétricas. Por ello, cuando no es posible determinar de manera objetiva cuál de los dos factores es el explicativo o el explicado, se debe optar por la utilización de la versión simétrica de estas medidas, cuyo valor es:

$$\lambda = \frac{\sum_{i=1}^h \max_j n_{ij} + \sum_{j=1}^k \max_i n_{ij} - \max_i n_{i.} - \max_j n_{.j}}{2N - \max_i n_{i.} - \max_j n_{ij}}$$

El valor de λ está comprendido entre λ_x y λ_y y presenta como inconveniente su gran sensibilidad a la presencia de totales marginales desequilibrados. Si λ se aproxima a 1 existe asociación entre X e Y , y si se aproxima a cero existirá independencia.

Existe también una serie de medidas utilizadas en los casos en que los atributos de la tabla de contingencia presentan sus modalidades ordenadas o son susceptibles de ordenación, siguiendo un orden natural.

Estas medias permiten, además de graduar el nivel de asociación entre los atributos, indicar la dirección de dicha asociación, según que la medida sea positiva o negativa. Suele haber tres casos extremos, como son la perfecta asociación positiva, la perfecta asociación negativa y la independencia. Hay que tener presente que antes de calcular el valor de los estadísticos que definen las medidas es necesario realizar la ordenación. Las medidas de este tipo más usualmente utilizadas son las siguientes:

- ***El coeficiente de correlación por rangos de Spearman:*** Ya fue estudiado anteriormente en este mismo capítulo.
- ***La Gamma de Goodman y Kruskall:*** Se basa en la relación relativa que siguen los rangos de dos atributos expresados en escala ordinal, es decir, hace referencia a la concordancia o discordancia entre los rangos de los atributos para los individuos observados. Su valor viene definido por la expresión $\gamma = S/(P+Q) = (P-Q)/(P+Q)$, donde P es el número de pares concordantes de individuos, es decir, pares de observaciones en los que los rangos de ambos factores siguen idéntica dirección (ambos crecen o ambos decrecen), y Q es el número de pares discordantes, es decir, pares de observaciones en los que los rangos de ambos factores siguen dirección opuesta (uno crece y otro decrece). La perfecta asociación positiva se da cuando $\gamma = 1$, La perfecta asociación negativa se da cuando $\gamma = -1$ y la independencia se da cuando $\gamma = 0$. La asociación será tanto mayor cuanto más se aproxime γ , en valor absoluto, a la unidad ($-1 \leq \gamma \leq 1$).

$$\gamma = \frac{\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} - \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs}}{\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} + \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs}}$$

- ***Coeficientes de correlación por rangos de Kendall τ_a , τ_b y τ_c :*** Se trata de medidas de asociación para factores ordinales. La más sencilla es τ_a . Para la obtención de esta medida se consideran los rangos de las N categorías de los atributos A y B , siendo (x_i, y_i) los pares de valores de los rangos. Se toman los rangos del primer atributo por orden de menor a mayor y se clasifican las observaciones de acuerdo con esos rangos ($x_i^*=1,2,\dots,N$), con lo que obtendremos una nueva secuencia de rangos para el segundo atributo (y_i^*), de tal forma que los pares (x_i^*, y_i^*) tienen las mismas componentes que los pares (x_i, y_i) para todo $i = 1, \dots, N$, cambiando sencillamente el orden de aparición. Se compara cada y_i^* con cada uno de los siguientes y se considera el indicador I_{ij} que vale -1 si en la comparación se ha producido un inversión del orden natural, y que vale 1 en caso contrario. Sumando todos los valores tomados por el indicador se obtiene un valor S . Entonces tenemos que $\tau_a = 2S/(N(N-1))$. La perfecta asociación positiva se da cuando $\tau_a = 1$, La perfecta asociación negativa se da cuando $\tau_a = -1$ y la independencia se da cuando $\tau_a = 0$. La asociación será tanto mayor cuanto más se aproxime τ_a , en valor absoluto, a la unidad. La

aplicación de esta medida exige que ambos factores posean el mismo número de categorías y que los totales marginales sean iguales. Se tiene que ($-1 \leq \tau_a \leq 1$). Las medidas τ_b y τ_c toman respectivamente los valores siguientes:

$$\tau_b = \frac{2S}{\sqrt{(P+Q+X_0)(P+Q+Y_0)}} \quad y \quad \tau_c = \frac{2mS}{N^2(m-1)}$$

donde tenemos que $m = \min(h, k)$, X_0 = número de pares ligados sobre el atributo X , Y_0 = número de pares ligados sobre el atributo Y , P , Q y S tienen el significado que el atribuido en la Gamma de Goodman y Kruskall. Un par es ligado si sus dos modalidades tienen algo en común en virtud de una determinada condición. El número de pares ligados sobre el atributo X (por filas) se obtiene multiplicando la frecuencia observada en cada celda de la tabla por las de cada una de las celdas situadas en la misma fila, pero a la derecha de ella, y sumando los productos. El número de pares ligados sobre el atributo Y (por columnas) se obtiene multiplicando la frecuencia observada en cada celda de la tabla por las de cada una de las celdas situadas en la misma columna, pero por debajo de ella, y sumando los productos. Se tiene $-1 \leq \tau_b \leq 1$ y $-1 \leq \tau_c \leq 1$ y τ_b sólo alcanza los valores extremos cuando la tabla es cuadrada. La perfecta asociación positiva se da cuando $\tau_b = 1$, $\tau_c = 1$. La perfecta asociación negativa se da cuando $\tau_b = -1$, $\tau_c = -1$, y la independencia se da cuando $\tau_b = 0$, $\tau_c = 0$. La asociación será tanto mayor cuanto más se aproximen τ_b , τ_c en valor absoluto a la unidad. También podemos expresar τ_b y τ_c así:

$$\tau_b = \frac{2(\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} - \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs})}{\sqrt{(N^2 - \sum_{i=1}^h (\sum_{j=1}^k n_{ij})^2)(N^2 - \sum_{j=1}^k (\sum_{i=1}^h n_{ij})^2)}}$$

$$\tau_c = \frac{2m(\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} - \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs})}{(m-1)N^2}$$

- **La d de Somers:** Si el atributo explicativo es X , e Y es el explicado, la medida propuesta por Somers es $d_{yx} = (P-Q)/(P+Q+Y_0)$. Si el atributo explicativo es Y y X es el explicado la medida propuesta por Somers es $d_{xy} = (P-Q)/(P+Q+X_0)$. Si no se distingue el atributo explicativo del explicado se utiliza la medida $d = (P-Q)/(P + Q + (X_0 + Y_0)/2)$. Se cumple que $-1 \leq d \leq 1$ y además $\tau_b^2 = d_{yx}d_{xy}$. La perfecta asociación positiva se da cuando $d = 1$, La perfecta asociación negativa se da cuando $d = -1$ y la independencia se da cuando $d = 0$. La asociación será tanto mayor cuanto más se aproxime d , en valor absoluto, a la unidad. Estas medidas también pueden expresarse como sigue:

$$d_{xy} = \frac{2(\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} - \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs})}{N^2 - \sum_{i=1}^h (\sum_{j=1}^k n_{ij})^2}$$

$$d_{yx} = \frac{2(\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} - \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs})}{N^2 - \sum_{j=1}^k (\sum_{i=1}^h n_{ij})^2}$$

$$d = \frac{4(\sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s>j}^k n_{rs} - \sum_{i=1}^h \sum_{j=1}^k n_{ij} \sum_{r>i}^h \sum_{s<j}^k n_{rs})}{(N^2 - \sum_{i=1}^h (\sum_{j=1}^k n_{ij})^2)(N^2 - \sum_{j=1}^k (\sum_{i=1}^h n_{ij})^2)}$$

La medida Eta: Se utiliza cuando el atributo dependiente se mide en un intervalo y el atributo independiente se mide en una escala ordinal o nominal. Cuando las filas son el atributo dependiente, el valor de la medida Eta es:

$$ER = \sqrt{1 - \frac{\sum_{j=1}^k (\sum_{i=1}^h x_i^2 n_{ij} - \frac{(\sum_{i=1}^h x_i^2 n_{ij})^2}{\sum_{i=1}^h n_{ij}})}{\sum_{j=1}^k \sum_{i=1}^h x_i^2 n_{ij} - \frac{(\sum_{i=1}^h \sum_{j=1}^k x_i n_{ij})^2}{N}}}$$

Cuando las filas son el atributo dependiente, el valor de la medida Eta es:

$$EC = \sqrt{1 - \frac{\sum_{i=1}^h (\sum_{j=1}^k y_j^2 n_{ij} - \frac{(\sum_{j=1}^k y_j^2 n_{ij})^2}{\sum_{j=1}^k n_{ij}})}{\sum_{j=1}^k \sum_{i=1}^h y_j^2 n_{ij} - \frac{(\sum_{i=1}^h \sum_{j=1}^k y_j n_{ij})^2}{N}}}$$

- **Coeficiente de incertidumbre**: Nos da el grado de relación lineal entre los dos atributos. Cuando las filas son el atributo dependiente su valor es $UR = (U(R) + U(C) - U(RC))/U(R)$. Cuando las columnas son el atributo dependiente su valor es $UC = (U(R) + U(C) - U(RC))/U(C)$. Cuando no se conoce la relación de dependencia entre atributos su valor es $U = (U(R) + U(C) - U(RC))/(U(R) + U(C))$. Los valores de $U(R)$, $U(C)$ y $U(RC)$ son los siguientes:

$$U(R) = -\sum_{i=1}^h \frac{\sum_{j=1}^k n_{ij}}{N} \ln\left(\frac{\sum_{j=1}^k n_{ij}}{N}\right) \quad U(C) = -\sum_{j=1}^k \frac{\sum_{i=1}^h n_{ij}}{N} \ln\left(\frac{\sum_{i=1}^h n_{ij}}{N}\right)$$

$$U(RC) = -\sum_{i=1}^h \sum_{j=1}^k \frac{n_{ij}}{N} \ln\left(\frac{n_{ij}}{N}\right)$$

El coeficiente R de Pearson: También nos da el grado de relación lineal entre los dos atributos, pero solo es bueno en un entorno de su valor 1, es decir, sólo es bueno para asegurar la asociación lineal positiva entre los atributos . Su expresión es:

$$R = \frac{\sum_{j=1}^k \sum_{i=1}^h x_i n_{ij} (\sum_{j=1}^k \sum_{i=1}^h y_j n_{ij})}{N}$$

$$R = \frac{\sum_{j=1}^k \sum_{i=1}^h x_i y_j n_{ij} - \sqrt{\left(\sum_{j=1}^k \sum_{i=1}^h x_i^2 n_{ij} - \frac{(\sum_{j=1}^k \sum_{i=1}^h x_i n_{ij})^2}{N} \right) \left(\sum_{j=1}^k \sum_{i=1}^h y_j^2 n_{ij} - \frac{(\sum_{j=1}^k \sum_{i=1}^h y_j n_{ij})^2}{N} \right)}}{\sqrt{\left(\sum_{j=1}^k \sum_{i=1}^h x_i^2 n_{ij} - \frac{(\sum_{j=1}^k \sum_{i=1}^h x_i n_{ij})^2}{N} \right) \left(\sum_{j=1}^k \sum_{i=1}^h y_j^2 n_{ij} - \frac{(\sum_{j=1}^k \sum_{i=1}^h y_j n_{ij})^2}{N} \right)}}$$

EL MODELO LOGARÍTMICO LINEAL

La finalidad principal en el análisis de las tablas de contingencia no es otra que constatar la existencia de la independencia de los factores objeto de estudio, y que integran la mencionada tabla. El problema se resuelve, consecutivamente, en dos pasos: En primer lugar, se procede a la contrastación de la hipótesis de independencia y después, si la hipótesis ha sido rechazada (no aparecen evidencias de independencia entre las variables), se realiza una estimación de la cuantificación del grado de asociación que presentan los factores.

Los problemas que resuelven las tablas de contingencia, siendo importantes, dejan sin respuesta cuestiones de gran importancia: Cuantificación (estimación) de la influencia individual que cada factor ejerce sobre las frecuencias, a través de sus diferentes niveles, influencia correspondiente a la acción conjunta de varios factores sobre la magnitud de las frecuencias de las celdas, en el conjunto de la tabla, en el caso que la contrastación de la independencia no haya conducido a rechazar la hipótesis. Estos temas son el objetivo de los modelos logarítmico lineales.

Para introducirnos en el análisis teórico de los modelos logarítmico lineales empezaremos, por simplicidad, con los que tienen como fin explicar la estructura de tablas de dos dimensiones ($R \times C$).

El fundamento de los modelos logarítmico lineales radica en el principio de independencia de dos variables aleatorias. Si dos variables aleatorias son independientes basta conocer el comportamiento individual de cada una (distribuciones marginales) para conocer el conjunto, y se verifica que la probabilidad conjunta p_{ij} se expresa, para todo i y para todo j , como $p_{ij} = p_{i\cdot}p_{\cdot j}$.

La condición de independencia se puede traducir en que, la probabilidad de que un elemento o individuo, presente simultáneamente las dos características (factores) en los niveles i -ésimo y j -ésimo, es igual al producto de la probabilidad de que presente la característica A en el nivel i -ésimo y la B en el j -ésimo; en otras palabras, la probabilidad conjunta (p_{ij}) es igual al producto de las probabilidades marginales ($p_{i\cdot}$ y $p_{\cdot j}$); o sea, el comportamiento conjunto de los dos factores se explica a través de los respectivos comportamientos individuales (marginales).

Si, por el contrario, consideramos la no existencia de independencia podemos expresar la probabilidad conjunta p_{ij} se expresa, para todo i y para todo j , como:

$$p_{ij} = p_{i\cdot}p_{\cdot j} k_{ij} \quad k_{ij} > 0$$

siendo k_{ij} la cuantificación del efecto conjunto del nivel i -ésimo del factor A y del j -ésimo del factor B, efecto conjunto al que daremos el nombre de **interacción de ambas variables o factores**.

Postular o aceptar la hipótesis de independencia de los dos factores, supone atribuir al elemento k_{ij} el valor unidad, para todo i y todo j .

Si tomamos logaritmos neperianos, se llega a que, en el caso de independencia

$$\ln p_{ij} = \ln p_{i\cdot} + \ln p_{\cdot j}$$

y cuando la independencia no existe

$$\ln p_{ij} = \ln p_{i\cdot} + \ln p_{\cdot j} + \ln k_{ij}$$

Estas dos expresiones constituyen una primera formulación de los modelos logarítmico lineales, para el caso de dos factores, es decir, cuando la tabla de contingencia es de dos dimensiones.

Aunque los modelos iniciales son, respectivamente, equivalentes a los logarítmicos, parece más razonable utilizar estas últimas formulaciones, modelos aditivos, en donde el término que discrimina la existencia o no de independencia es ahora $\ln k_{ij}$, que toma el valor cero cuando estamos en presencia de independencia, valor más acorde con la idea de ausencia de asociación que se pretende representar, en vez del valor unidad que se le asignaba al parámetro k_{ij} en el modelo multiplicativo inicial para este mismo caso.

Mediante estos modelos pretendemos expresar, en términos aditivos y no multiplicativos, que la mayor o menor presencia de los individuos en una celda determinada (en la población, esta presencia se mide a través de la probabilidad p_{ij}) depende de la estructura de las distribuciones marginales y, en el caso general, de la existencia de interacciones, correspondientes a la acción conjunta de los dos factores o características.

Efectos principales

En vez de centrar nuestra atención en las probabilidades de cada celda, como es el caso de los modelos, utilizaremos, como base de las estimaciones, las frecuencias observadas, dado que ésta es la información de que disponemos. Si recordamos que las frecuencias esperadas (E_{ij}) se calculan, en una tabla con un número total de elementos igual a N , como el producto de la probabilidad de cada celda por este número total, $E_{ij} = Np_{ij}$, y que en el caso de independencia o ausencia de asociación resulta

$$E_{ij} = Np_{ij} = Np_i \cdot p_j$$

y dado que como sabemos,

$$E_{i \cdot} = Np_{i \cdot} \text{ y } E_{\cdot j} = Np_{\cdot j}$$

Se puede escribir:

$$E_{ij} = E_{i \cdot} E_{\cdot j} / N \Rightarrow \ln E_{ij} = -\ln N + \ln E_{i \cdot} + \ln E_{\cdot j}$$

Este modelo tiene semejanza formal con el del análisis de la varianza, por lo cual es aconsejable la utilización de la terminología de esta importante técnica de análisis estadístico.

A continuación efectuamos una serie de transformaciones a fin de hacer operativo el modelo. El factor A se presenta con r niveles: $i = 1 \dots r$, y el factor B con c niveles: $j = 1 \dots c$.

Sumando los dos miembros de la ecuación anterior con respecto a i (factor A), tenemos:

$$\sum_{i=1}^r \ln E_{il} = -r \ln N + \sum_{i=1}^r \ln E_{i \cdot} + r \ln E_{\cdot j}$$

A continuación si calculamos la suma con respecto a j (factor B) tenemos:

$$\sum_{j=1}^c \ln E_{ij} = -c \ln N + c \ln E_{i \cdot} + \sum_{j=1}^c \ln E_{\cdot j}$$

Por último, si la suma se realiza respecto a i y j (los dos factores), tenemos:

$$\sum_{i=1}^r \sum_{j=1}^c \ln E_{ij} = -rc \ln N + c \sum_{i=1}^r \ln E_{i\cdot} + r \sum_{j=1}^c \ln E_{\cdot j}$$

Dividiendo las tres últimas expresiones por r , por c , y por el producto rc respectivamente, obtenemos:

$$\frac{1}{r} \sum_{i=1}^r \ln E_{ij} = -\ln N + \frac{1}{r} \sum_{i=1}^r \ln E_{i\cdot} + \ln E_{\cdot j}$$

$$\frac{1}{c} \sum_{j=1}^c \ln E_{ij} = -\ln N + \ln E_{i\cdot} + \frac{1}{c} \sum_{j=1}^c \ln E_{\cdot j}$$

$$\frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \ln E_{ij} = -\ln N + \frac{1}{r} \sum_{i=1}^r \ln E_{i\cdot} + \frac{1}{c} \sum_{j=1}^c \ln E_{\cdot j}$$

Restando a la última expresión la suma de las dos primeras, se tiene:

$$\begin{aligned} -\ln N + \ln E_{i\cdot} + \ln E_{\cdot j} &= \frac{1}{r} \sum_{i=1}^r \ln E_{ij} + \frac{1}{c} \sum_{j=1}^c \ln E_{ij} - u = \\ &= u + \left(\frac{1}{c} \sum_{j=1}^c \ln E_{ij} - u \right) + \left(\frac{1}{r} \sum_{i=1}^r \ln E_{ij} - u \right) = u + u_{1(i)} + u_{2(j)} \end{aligned}$$

dónde hemos definido los parámetros:

$$u = \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \ln E_{ij}$$

$$u_{1(i)} = \frac{1}{c} \sum_{j=1}^c \ln E_{ij} - \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \ln E_{ij}$$

$$u_{2(j)} = \frac{1}{r} \sum_{i=1}^r \ln E_{ij} - \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \ln E_{ij}$$

resultado que ya podemos escribir la expresión:

$$\ln E_{ij} = u + u_{1(i)} + u_{2(j)}$$

que nos dice que el logaritmo neperiano de la frecuencia esperada puede descomponerse en la suma de tres parámetros, lo que pone de manifiesto la similitud con el modelo del análisis de la varianza a que antes aludíamos, ya que, como vamos a ver a continuación, cada uno de estos tres parámetros va a tener un significado similar al que poseen los del referido modelo del análisis de la varianza.

El sumando u cuantifica el valor que adoptarían los logaritmos de las frecuencias esperadas si los dos factores no ejercieran ningún efecto. El sumando $u_{1(i)}$ mide el efecto que produce el nivel i -ésimo de la variable A, factor A, o más claramente la influencia que la fila i -ésima ejerce sobre el logaritmo del número de elementos o individuos que posean ese nivel de A. De manera análoga, el sumando $u_{2(j)}$ evalúa el efecto que el nivel j -ésimo del factor B, efecto que la columna j -ésima, ejerce sobre la aparición de elementos en ese nivel.

Los parámetros u_1 y u_2 reciben el nombre de efectos principales o directos. Los valores positivos de los efectos directos (principales) indican que el nivel en cuestión actúa favoreciendo la presencia de individuos en esa fila o columna; dicho con otras palabras, la probabilidad de aparición de individuos en esa fila tiende a ser «alta».

Los valores negativos indican la situación contraria, esto es, el nivel no favorece, penaliza, la presencia de individuos en esa situación.

La primera relación del principio de esta página, no es otra cosa que la media general de los logaritmos de las $r \times c$ frecuencias estimadas (recordemos que en una tabla $R \times C$ hay $r \times c$ celdas). La segunda relación es igual a la diferencia de dos términos de los cuales el primero es la media de la distribución marginal, respecto al factor B (columnas), de los logaritmos de las frecuencias esperadas y el segundo término es la media general u . Esta diferencia nos dice que los efectos producidos por los distintos niveles del factor A se miden a través de las desviaciones de las medias marginales de cada nivel respecto a la media general: se ha eliminado el efecto general que existía si todas las celdas tuvieran el mismo número de elementos.

Si sumamos los r términos $u_{1(i)}$ llegamos, fácilmente, a que la suma es igual a cero.

$$\sum_{i=1}^r u_{1(i)} = 0$$

Esta conclusión, que coincide plenamente con lo que sucede en el análisis de la varianza, tiene una consecuencia importante como es que los efectos producidos por los niveles de una característica, tal y como han sido definidos, no son independientes entre sí, pues su suma es igual a cero, lo que implica la relatividad del concepto «efecto», no siendo posible separarlo del contexto de la estructura concreta de la tabla que se analiza, puesto que cualquier modificación de dicha estructura conduce a una modificación, cuantitativa y quizás de signo, de las estimaciones de los efectos.

Los comentarios referentes a la tercera relación son, por simetría, exactamente iguales que los realizados con referencia a la segunda; sin más que efectuar las trasposiciones pertinentes. En particular, los efectos del factor B deben cumplir, análogamente a los de A, la condición:

$$\sum_{j=1}^c u_{2(j)} = 0$$

Interacciones

Hasta ahora, para exponer las similitudes que existen entre los modelos lineales y el modelo del análisis de la varianza, hemos estudiado el modelo logarítmico lineal más simple, suponiendo la existencia de independencia, lo que implica la condición $p_{ij} = p_i \cdot p_j$. Es inmediata la extensión de este modelo simple al más complicado por admitir la falta de independencia, falta que se materializa en la presencia del término de las interacciones, como expresa el modelo $\ln p_{ij} = \ln p_i + \ln p_j + \ln k_{ij}$, que implica la condición $p_{ij} = p_i \cdot p_j \cdot k_{ij}$, cuyos cálculos omitimos por considerarlos no necesarios en toda esta exposición.

El modelo logarítmico lineal sin independencia, con presencia de interacciones por lo tanto, será

$$\ln E_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

Los términos primero, segundo y tercero, del segundo miembro, tienen las mismas interpretaciones que anteriormente. El tercer sumando $u_{12(ij)}$ expresa la cuantificación de la acción conjunta del nivel j-ésimo del factor A, o efecto de la fila i , y del nivel j-ésimo del factor B, o efecto de la columna j (en términos del análisis de la varianza la interacción de las dos variables o factores) de tal forma que si tiene lugar la interacción su existencia influye en las diferencias de las frecuencias entre celdas, supuesto eliminados los efectos de filas y columnas [$u_{1(i)}$ y $u_{2(j)}$] y media general (u).

De la misma manera que en los efectos principales existen unas condiciones que deben cumplir los efectos de los niveles de cada factor. En el caso de la interacción aparecen dos condiciones (restricciones) que verifican los efectos de las interacciones (por la forma en que han sido definidos) y son las siguientes:

$$\sum_{i=1}^r u_{12(ij)} = 0 \quad \text{y} \quad \sum_{j=1}^c u_{12(ij)} = 0$$

Cada una de las dos sumas marginales (respecto a las filas y a las columnas) son iguales a cero, lo que supone la no independencia de los efectos de las interacciones, y su consiguiente relatividad, en línea total con lo que sucedía con los efectos principales.

El modelo especificado para dos factores es fácilmente generalizable para cualquier número de ellos. A continuación vamos a hacerlo para tres, lo que además nos dará pie para acercarnos a la problemática de los modelos logarítmico lineales.

Tres factores o características, cuando los elementos o individuos que los poseen se clasifican con respecto a ellos, dan lugar a una tabla de contingencia de tres dimensiones. De forma análoga el caso de dos factores hemos de tener en cuenta una serie de efectos que contribuyen a la tendencia de un elemento a pertenecer a una u otra celda de la tabla. Estos efectos son: tres efectos directos o principales, uno por cada factor (u_1 , u_2 y u_3); a continuación tenemos los efectos conjuntos de cada par de características, denominadas interacciones de orden uno (u_{12} , u_{13} y u_{23}); por último el efecto conjunto de las tres características, interacciones de orden dos (u_{123}).

Podemos expresar todo esto mediante el modelo:

$$\ln E_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12}(ij) + u_{13}(ik) + u_{23}(jk) + u_{123}(ijk)$$

Modelo saturado

En total, el número de parámetros del modelo $\ln E_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12}(ij)$ será: $1 + (r - 1) + (c - 1) + (r - 1)(c - 1) = rc$

Este número resulta igual al de celdas de la tabla ($R \times C$). De manera similar podríamos determinar los parámetros que figuran en los modelos de 3, 4 o más dimensiones.

A partir de todo esto definimos un modelo saturado si contiene en su formulación tantos parámetros independientes como celdas tiene la tabla a la que es aplicado; dicho de otra forma más operativa, un modelo es saturado si incluye todos los posibles efectos principales e interacciones. En caso contrario, si falta algún o algunos parámetros, se le denomina no saturado.

Por otra parte, el modelo saturado reproduce exactamente las frecuencias observadas.

Modelo jerárquico

El concepto de modelo jerárquico se refiere a la «coherencia» interna de los modelos no saturados: si en un modelo falta un término (se supone que se ha establecido sobre ese término la hipótesis de que es igual a cero) para considerarle como jerárquico, deberán estar excluidos todos los parámetros de orden superior que contengan la combinación de subíndices fijos del que no aparece. Esto se verá con más facilidad con un ejemplo: si en el modelo de cuatro factores se suprime el término $u_{23(jk)}$, para que el modelo resultante sea jerárquico, deberán estar ausentes los términos que contengan conjuntamente la combinación de subíndices fijos 2 y 3.

INDEPENDENCIA Y ASOCIACIÓN EN MODELOS LOGARÍTMICO LINEALES

Para ilustrar los conceptos de independencia y asociación utilizaremos el modelo saturado de tres dimensiones (RxCxS):

$$\ln E_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12}(ij) + u_{13}(ik) + u_{23}(jk) + u_{123}(ijk)$$

Tenemos los siguientes conceptos:

Independencia total, completa o global

Los tres factores son total o completamente independientes si lo son tanto simultáneamente como dos a dos, lo que conduce a que, para todo i, j y k, todas las interacciones sean nulas, es decir: $u_{12}=u_{13}=u_{23}=u_{123}=0$

Independencia parcial: un factor completamente independiente de los demás

Recurriendo al modelo tridimensional diremos que el factor 1 es completamente independiente de los otros dos factores (2, 3) cuando todas las interacciones en las que aparece ese factor son nulas: $u_{12}=u_{13}=u_{123}=0$

Independencia condicional

Siguiendo con el ejemplo del modelo de tres factores, podemos decir que los factores 1 y 2 presentan independencia condicional cuando son independientes entre sí, para cada nivel del tercer factor, pudiendo estar ambos factores asociados a su vez con este tercer factor, teniendo entonces que: $u_{12}=u_{123}=0$

Asociación parcial

El modelo presenta asociación parcial entre los tres factores cuando las interacciones de segundo orden son nulas, $u_{123}=0$ siendo diferentes de cero todas las interacciones de primer orden.

ESTIMACIÓN MÁXIMO VERO SIMIL DE LOS PARÁMETROS DEL MODELO

Tablas de contingencia bidimensionales

Comenzaremos exponiendo con detalle el método directo y exacto en tablas de contingencia (R X C) y sin ceros estructurales.

Llamamos n_{ij} a la frecuencia observada correspondiente a la combinación del nivel i-ésimo del factor A y del j-ésimo del factor B: $n_{i\cdot}$ y $n_{\cdot j}$ son las frecuencias marginales de cada uno de los dos factores, siendo n el número total de elementos que han sido clasificados en la tabla de contingencia; su cálculo, recordamos, es el siguiente:

$$n_{i\cdot} = \sum_{j=1}^c n_{ij}; \quad n_{\cdot j} = \sum_{i=1}^r n_{ij}; \quad N = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$$

Partimos del modelo saturado de la tabla (R X C)

$$\ln E_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

y establecemos la hipótesis de independencia, es decir que para todo i y j las interacciones son nulas,

$$u_{12(ij)} = 0$$

Llamamos \hat{E}_{ij} a la estimación de la frecuencia esperada E_{ij} . Las estimaciones se realizan por el método de la máxima verosimilitud y la solución general es:

$$\hat{E}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$$

Una vez conseguidas las estimaciones de las frecuencias esperadas estamos en condiciones de contrastar la hipótesis de independencia.

El principal problema reside en la necesidad de disponer de un estadístico que cuantifique en qué medida un modelo concreto se ajusta adecuadamente a la estructura de una tabla de contingencia concreta, o lo que es análogo, cuanto explica el modelo la estructura de la tabla. Las dos medidas más usuales son el estadístico X^2 de Pearson y la razón de verosimilitud (G^2). La primera de las dos tiene por expresión, como se ha indicado anteriormente,

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

siendo la expresión de la segunda medida

$$G^2 = 2 \sum_{i=1}^r \sum_{j=1}^c n_{ij} \ln \frac{n_{ij}}{\hat{E}_{ij}}$$

Estos dos estadísticos, para la situación que la hipótesis cierta sea la de independencia, se distribuyen como una ji-cuadrado con $(r-1)(c-1)$ grados de libertad.

Si el número total de elementos de la tabla de contingencia, N , es elevado los dos tests son equivalentes. La ventaja del primero sobre el segundo radica en el hecho de la posible partición de X^2 en tantos sumandos como grados de libertad tenga.

La generalización de las expresiones 2 y 3 a una tabla multidimensional es inmediata. En las dos fórmulas, n_{ij} representa las frecuencias observadas, las que componen la tabla de contingencia, y \hat{E}_{ij} las frecuencias esperadas en cada celda bajo las diferentes hipótesis que se establezcan en cada modelo logarítmico lineal. En la práctica estas frecuencias esperadas se sustituyen por sus estimaciones máximo verosímiles, y es en este proceso de estimación donde radica la menor o mayor complejidad del cálculo y, por consiguiente, la solución del modelo.

Para obtener los valores concretos de X^2 y verosimilitud (G^2) sustituiremos en ambas expresiones, 2 y 3, la estimación \hat{E}_{ij} obtenida en 1, resultando:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(n_{ij} - \frac{n_{i \cdot} n_{\cdot j}}{N} \right)^2}{\frac{n_{i \cdot} n_{\cdot j}}{N}} = \sum_{i=1}^r \sum_{j=1}^c \frac{(Nn_{ij} - n_{i \cdot} n_{\cdot j})^2}{Nn_{i \cdot} n_{\cdot j}}$$

y análogamente para el estadístico

$$G^2 = 2 \sum_{i=1}^r \sum_{j=1}^c n_{ij} \ln \frac{n_{ij}}{\hat{E}_{ij}} = 2 \sum_{i=1}^r \sum_{j=1}^c n_{ij} \ln \frac{n_{ij}}{\frac{n_{i \cdot} n_{\cdot j}}{N}} = 2 \sum_{i=1}^r \sum_{j=1}^c n_{ij} \ln \frac{Nn_{ij}}{n_{i \cdot} n_{\cdot j}}$$

Si el resultado de la contrastación, fijado el nivel de significación adecuado, ha sido aceptar la hipótesis de independencia $u_{12} = 0$ equivale, como sabemos, a aceptar que el modelo logarítmico lineal tiene por expresión:

$$\ln E_{ij} = u + u_{1(i)} + u_{2(j)}$$

El paso siguiente será la estimación de los efectos principales, debiendo tener presente que, al haber aceptado la hipótesis de independencia, las estimaciones de los efectos deberán ser calculados aplicando esta condición, expresión 1. Recurriendo a las fórmulas 12, 13, 14 del capítulo anterior, y sustituyendo en ellas E_{ij} por \hat{E}_{ij} en 1, llegamos a los siguientes resultados:

Estimación de la media general

$$\begin{aligned}\hat{u} &= \frac{\sum_{i=1}^r \sum_{j=1}^c \ln \hat{E}_{ij}}{rc} = \frac{\sum_{i=1}^r \sum_{j=1}^c \ln n_{i\cdot} + \sum_{i=1}^r \sum_{j=1}^c \ln n_{\cdot j} - \sum_{i=1}^r \sum_{j=1}^c \ln N}{rc} = \\ &= \frac{c \sum_{i=1}^r \ln n_{i\cdot} + r \sum_{j=1}^c \ln n_{\cdot j} - rc \ln N}{rc} = \frac{\sum_{i=1}^r \ln n_{i\cdot}}{r} + \frac{\sum_{j=1}^c \ln n_{\cdot j}}{c} - \ln N\end{aligned}$$

Estimación del efecto principal del factor A

$$\hat{u}_{1(i)} = \frac{\sum_{j=1}^c \ln \hat{E}_{ij}}{r} - \frac{\sum_{i=1}^r \sum_{j=1}^c \ln \hat{E}_{ij}}{rc} = \frac{c \ln n_{i\cdot} + \sum_{j=1}^c \ln n_{\cdot j} - c \ln N}{c} - \hat{u} = \ln n_{\cdot j} - \frac{1}{r} \sum_{j=1}^c \ln n_{\cdot j}$$

Estimación del efecto principal del factor B

$$\hat{u}_{2(j)} = \frac{\sum_{i=1}^r \ln \hat{E}_{ij}}{r} = \frac{\sum_{i=1}^r \sum_{j=1}^c \ln \hat{E}_{ij}}{rc} = \frac{r \ln n_{\cdot j} + \sum_{i=1}^r \ln n_{i\cdot} - r \ln N}{r} - \hat{u} = \ln n_{\cdot j} - \frac{1}{c} \sum_{i=1}^r \ln n_{i\cdot}$$

Se comprueba fácilmente las dos condiciones que deben cumplir los efectos.

$$\sum_{i=1}^r \hat{u}_{1(i)} = 0 \quad y \quad \sum_{j=1}^c \hat{u}_{2(j)} = 0$$

Si la contrastación conduce a no aceptar la existencia de independencia, es decir que no todas las interacciones son distintas de cero, el modelo saturado al que se ajustan las frecuencias, deberá ser:

$$\ln E_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

La no aceptación de la independencia de las dos características, o lo que es equivalente la presencia de asociación, lleva consigo que la expresión 1 no puede ser utilizada en los cálculos posteriores a la hora de obtener las estimaciones máximo verosímiles de los efectos, puesto que sólo es válida dicha expresión en el supuesto de independencia, por lo cual las estimaciones máximo verosímiles precisas se han de basar en el conocimiento que tenemos de la estructura poblacional de los dos factores, y este conocimiento no es otro que la propia tabla de contingencia, por lo cual la estimación máximo verosímil de las frecuencias esperadas será en este caso $\hat{E}_{ij} = n_{ij}$, que sustituido en 12, 13, y 14 del capítulo 18, proporciona las siguientes estimaciones de los parámetros

$$\begin{aligned}\hat{u} &= \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \ln n_{ij} \\ \hat{u}_{1(i)} &= \frac{1}{c} \sum_{j=1}^c \ln n_{ij} - \hat{u} \\ \hat{u}_{2(j)} &= \frac{1}{r} \sum_{i=1}^r \ln n_{ij} - \hat{u}\end{aligned}$$

Obtenidas las estimaciones de estos parámetros, el paso siguiente consiste en el cálculo de las interacciones $\hat{u}_{12(ij)}$, utilizando para ello la relación:

$$\hat{u}_{12(ij)} = \ln n_{ij} - [\hat{u} + \hat{u}_{1(i)} + \hat{u}_{2(j)}]$$

Las estimaciones $\hat{u}_{12(ij)}$ verifican, como se indicó líneas atrás, las dos condiciones:

$$\sum_{i=1}^r \hat{u}_{12(ij)} = 0 \quad y \quad \sum_{j=1}^c \hat{u}_{12(ij)} = 0$$

Efectuados los cálculos pertinentes es posible comparar entre sí el efecto que producen los diferentes niveles de cada factor sobre la cuantía de las frecuencias, así como, si no se ha aceptado la hipótesis de independencia, los distintos efectos que inducen conjuntamente las combinaciones de los niveles de las dos características.

ANÁLISIS DE LOS RESIDUOS

Los tests χ^2 y G^2 proporcionan una idea global de cómo los datos (frecuencias observadas) de la tabla de contingencia proceden de una población que obedece al modelo logarítmico lineal que hayamos considerado, es decir, permite establecer la bondad del ajuste de esos datos al modelo establecido, pero no informan sobre la relevancia particular en ese ajuste de los términos u del modelo, ni de los casos extremos que puedan darse en cada una de las celdas de la clasificación cruzada.

Dos tipos, pues, de análisis deben completar el diagnóstico que, sobre las estimaciones, proporcionan los conocidos tests χ^2 y G^2 en el estudio de una tabla, que se pretenda describir a través de un modelo logarítmico lineal. Estas dos clases de análisis se basan en la estimación de las diferencias entre las frecuencias observadas y las proporcionadas por el modelo en cuestión, diferencias denominadas residuos, y en su grado concreto de significación estadística. Veámos cuál es el modo de proceder.

Análisis de los residuos de los parámetros del modelo

Este método es estrictamente aplicable al caso de modelos saturados y completos, es decir, donde necesariamente todas las estimaciones de las frecuencias esperadas sean mayores que cero, en ausencia de ceros estructurales.

La significación estadística de los parámetros u de un modelo saturado puede estudiarse a través de su estimación tipificada, bajo el supuesto de que dicho parámetro u sea nulo, es decir, usando los residuos tipificados

$$\hat{w} = \frac{\hat{u} - 0}{\hat{S}_u}$$

donde la estimación del error típico \hat{S}_u también se obtiene de acuerdo al método de estimación directa o aproximada que se halla utilizado.

Para grandes muestras \hat{w} sigue una distribución asintóticamente normal con media cero y varianza unidad. Por tanto para, por ejemplo, un nivel de significación del 10% un valor de $|\hat{w}|$ mayor que 1,65 indica que dicho parámetro u es significativamente distinto de cero.

El uso de este análisis residual permite ordenar tanto los efectos principales e interacciones de acuerdo con su grado de significación y magnitud como, en su caso, reducir el modelo cuando algún parámetro u sea significativamente igual a cero, marcándonos la pauta del modelo no saturado que parezca más adecuado, así como la posibilidad de colapso de la tabla.

Goodrnan obtuvo una estimación asintótica del error típico S_u , igual para todos los términos u de un modelo saturado correspondiente a una tabla bidimensional ($R \times C$). Dicha estimación es:

$$\hat{S}_u = \frac{1}{(rc)^2} \sum_{i=1}^r \sum_{j=1}^c \frac{1}{n_{ij}}$$

de manera que, en este caso, los residuos de los parámetros son:

$$\hat{w}_1 = \frac{\hat{u}_1}{\hat{S}_u}, \quad \hat{w}_2 = \frac{\hat{u}_2}{\hat{S}_u} \quad \text{y} \quad \hat{w}_{12} = \frac{\hat{u}_{12}}{\hat{S}_u}$$

En modelos no saturados todas las expresiones anteriores, relativas al error típico de los parámetros u , pueden considerarse como cotas superiores de las estimaciones asintóticas de dichos errores, y los correspondientes tests pueden utilizarse también como cotas superiores de los correspondientes a estos modelos no saturados.

Análisis de los residuos en las celdas

Resulta de interés el estudio del grado de ajuste en cada una de las celdas de una tabla de contingencia, con objeto de verificar si, a pesar de que el ajuste global del modelo sea satisfactorio, existen algunas celdas donde no lo sea tanto, así como para detectar la presencia de valores extremos, o para establecer el patrón positivo o negativo de las desviaciones de las celdas.

Al calcular estas desviaciones entre los valores observados, n_{ij} , y las estimaciones de los esperados, \hat{E}_{ij} , quizá ocurra que se obtengan diferencias grandes, lo que indicará que las combinaciones de los niveles donde tal cosa suceda presentan particularidades de interés, o bien que los valores de las celdas son, como decíamos, valores extremos. Por ello, el análisis de tales desviaciones es capaz de arrojar, en muchas ocasiones, importante luz sobre la estructura de la tabla de contingencia.

Como casos límites una desviación puede ser alta simplemente por que lo sean los dos términos que la integran, o pequeña, por serlo también los dos términos; en cualquier circunstancia, estos valores límites pueden enmascarar situaciones extremas. En los dos casos el «elemento perturbador» es la propia magnitud, alta o baja, de las dos frecuencias (observada y esperada), por lo cual se hace preciso eliminar esta influencia.

Uno de los procedimientos de determinación de la significación de tales desviaciones se basa en el análisis de los *residuos tipificados*

$$d_{ij} = \frac{n_{ij} - \hat{E}_{ij}}{S_{\hat{E}_{ij}}}$$

en donde $S_{\hat{E}_{ij}}$ es el error típico de la estimación de las frecuencias esperadas E_{ij} . Y si tomamos como base de referencia del grado de bondad del ajuste el test χ^2 , los residuos tipificados son:

$$d_{ij} = \frac{n_{ij} - \hat{E}_{ij}}{\sqrt{\hat{E}_{ij}}}$$

La estimación del error típico implícitamente supone una distribución de Poisson para el número de observaciones de cada una de las celdas, por lo cual

$$S_{\hat{E}_{ij}} = \sqrt{\hat{E}_{ij}}$$

Estos residuos tipificados d_{ij} se distribuyen asintóticamente como una normal de media cero y varianza uno, por lo que, al igual que en el caso anterior de los residuos tipificados de los parámetros, un valor de d_{ij} mayor que, por ejemplo, 1,65 indica que la desviación $n_{ij} - \hat{E}_{ij}$ es significativamente distinta de cero, al nivel de significación del 10%.

Para aclarar los conceptos anteriores presentamos un ejemplo, continuación M primero de este capítulo.

MODELOS LOGARÍTMICO LINEALES Y TABLAS DE CONTINGENCIA CON SPSS

EL PROCEDIMIENTO TABLAS DE CONTINGENCIA

El procedimiento Tablas de contingencia crea tablas de clasificación doble y múltiple y, además, proporciona una serie de pruebas y medidas de asociación para las tablas de doble clasificación. La estructura de la tabla y el hecho de que las categorías estén ordenadas o no, determinan las pruebas o medidas que se utilizan.

Los estadísticos de tablas de contingencia y las medidas de asociación sólo se calculan para las tablas de doble clasificación. Si especifica una fila, una columna y un factor de capa (variable de control), el procedimiento Tablas de contingencia crea un panel de medidas y estadísticos asociados para cada valor del factor de capa (o una combinación de valores para dos o más variables de control). Por ejemplo, si GÉNERO es un factor de capa para una tabla de CASADO (sí, no) en función de VIDA (vida emocionante, rutinaria o aburrida), los resultados para una tabla de doble clasificación para las mujeres se calculan de forma independiente de los resultados de los hombres y se imprimen en paneles uno detrás del otro.

En cuanto a estadísticos, se obtiene chi-cuadrado de Pearson, chi-cuadrado de la razón de verosimilitud, prueba de asociación lineal por lineal, prueba exacta de Fisher, chi-cuadrado corregido de Yates, r de Pearson, rho de Spearman, coeficiente de contingencia, phi, V de Cramer, lambdas simétricas y asimétricas, tau de Kruskal y Goodman, coeficiente de incertidumbre, gamma, d de Somers, tau-b de Kendall, tau-c de Kendall, coeficiente eta, kappa de Cohen, estimación de riesgo relativo, razón de ventajas, prueba de McNemar, estadístico de Cochran y Mantel-Haenszel.

En cuanto a los datos, para definir las categorías de cada variable, utilice valores de una variable numérica o de cadena corta (ocho caracteres o menos). Por ejemplo, para GÉNERO, podría codificar los datos como 1 y 2 o como varón y mujer.

En algunos estadísticos y medidas se asume que hay unas categorías ordenadas (datos ordinales) o unos valores cuantitativos (datos de intervalos o de proporciones), como se explica en la sección sobre los estadísticos. Otros estadísticos son válidos cuando las variables de la tabla tienen categorías no ordenadas (datos nominales). Para los estadísticos basados en chi-cuadrado (phi, V de Cramer y coeficiente de contingencia), los datos deben ser una muestra aleatoria de una distribución multinomial. Las variables ordinales pueden ser códigos numéricos que representen categorías (p. ej., 1 = bajo, 2 = medio, 3 = alto) o valores de cadena. Sin embargo, se supone que el orden alfabético de los valores de cadena indica el orden correcto de las categorías. Por ejemplo, en una variable de cadena cuyos valores sean bajo, medio, alto, se interpretaría el orden de las categorías como alto, bajo, medio (orden que no es el correcto). Por norma general, se puede indicar que es más fiable utilizar códigos numéricos para representar datos ordinales.

Para obtener tablas de contingencia, elija en los menús *Analizar* → *Estadísticos descriptivos* → *Tablas de contingencia* (Figura 12-1) y seleccione una o más variables de fila y una o más variables de columna (Figura 12-2). Si lo desea, tiene la posibilidad de seleccionar una o más variables de control o capa, pulsar en *Estadísticos* (Figura 12-3) para obtener pruebas y medidas de asociación para tablas o subtablas de doble clasificación, pulsar en *Casillas* (Figura 12-4) para obtener porcentajes, residuos y valores esperados y observados o pulsar en *Formato* para controlar el orden de las categorías (ascendente o descendente).

Cruzaremos el sexo (*sexo*) y la categoría laboral (*catlab*) en el fichero EMPLEADOS. Pulsando *Continuar* en cada Figura se aceptan sus especificaciones y al pulsar *Aceptar* en la Figura 12-2, se obtiene la salida del procedimiento (Figuras 12-5 a 12-11).

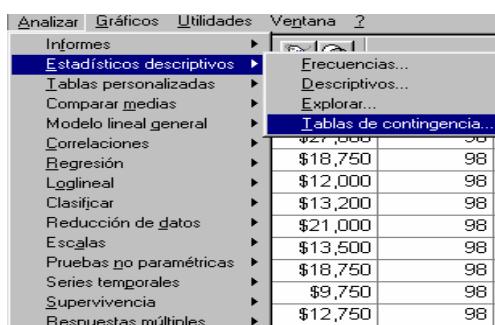


Figura 12-1

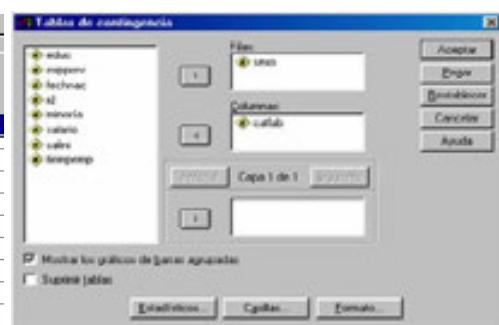


Figura 12-2

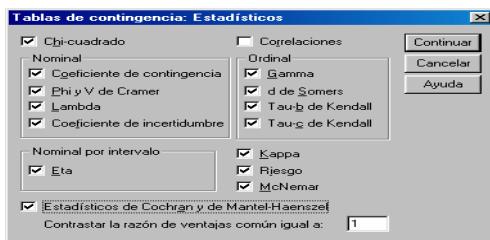


Figura 12-3



Figura 12-4

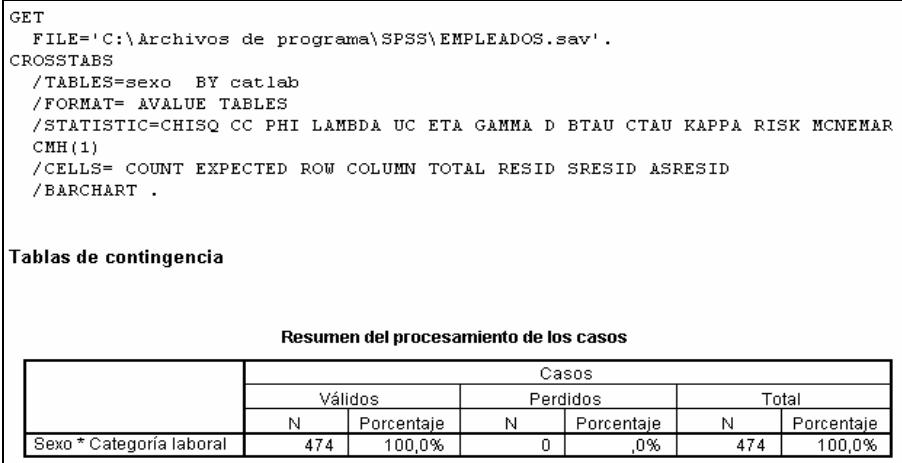


Figura 12-5

			Tabla de contingencia Sexo * Categoría laboral				
			Categoría laboral		Total		
Sexo	Hombre	Administrativo	Seguridad	Directivo			
		157	27	74	258		
		Frecuencia esperada	197,6	14,7	45,7	258,0	
		% de Sexo	60,9%	10,5%	28,7%	100,0%	
		% de Categoría laboral	43,3%	100,0%	88,1%	54,4%	
		% del total	33,1%	5,7%	15,6%	54,4%	
		Residuo	-40,6	12,3	28,3		
		Residuos tipificados	-2,9	3,2	4,2		
		Residuos corregidos	-8,8	4,9	6,8		
Sexo	Mujer	Recuento	206	0	10	216	
		Frecuencia esperada	165,4	12,3	38,3	216,0	
		% de Sexo	95,4%	,0%	4,6%	100,0%	
		% de Categoría laboral	56,7%	,0%	11,9%	45,6%	
		% del total	43,5%	,0%	2,1%	45,6%	
		Residuo	40,6	-12,3	-28,3		
		Residuos tipificados	3,2	-3,5	-4,6		
		Residuos corregidos	8,8	-4,9	-6,8		
		Recuento	363	27	84	474	
Total			Frecuencia esperada	363,0	27,0	84,0	
			% de Sexo	76,6%	5,7%	17,7%	
			% de Categoría laboral	100,0%	100,0%	100,0%	
			% del total	76,6%	5,7%	17,7%	
						100,0%	

Figura 12-6

Medidas direccionales						
			Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,150	,054	2,590	,010
		Sexo dependiente	,227	,078	2,590	,010
		Categoría laboral dependiente	,000	,000	,	,
	Tau de Goodman y Kruskal	Sexo dependiente	,167	,024		,000 ^d
		Categoría laboral dependiente	,123	,021		,000 ^d
		Coeficiente de incertidumbre	,148	,022	6,237	,000 ^e
Ordinal por ordinal	d de Somer	Simétrica	,146	,023	6,237	,000 ^e
		Sexo dependiente	,146	,023	6,237	,000 ^e
		Categoría laboral dependiente	,149	,022	6,237	,000 ^e
Nominal por intervalo	Eta	Simétrica	,386	,033	-9,999	,000
		Sexo dependiente	,446	,037	-9,999	,000
		Categoría laboral dependiente	,340	,034	-9,999	,000
Nominal por nominal	Eta	Sexo dependiente	,409			
		Categoría laboral dependiente	,378			

^a. Asumiendo la hipótesis alternativa.
^b. Empleando el error típico asintótico basado en la hipótesis nula.
^c. No se puede efectuar el cálculo porque el error típico asintótico es igual a cero.
^d. Basado en la aproximación chi-cuadrado.
^e. Probabilidad del chi-cuadrado de la razón de verosimilitud.

Figura 12-7

Medidas simétricas						
			Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Phi		,409			,000
		V de Cramer	,409			,000
		Coeficiente de contingencia	,379			,000
	Ordinal por ordinal	Tau-b de Kendall	-,389	,033	-9,999	,000
		Tau-c de Kendall	-,338	,034	-9,999	,000
		Gamma	-,837	,051	-9,999	,000
Medida de acuerdo N de casos válidos	Kappa		, ^c			
		N de casos válidos	474			

^a. Asumiendo la hipótesis alternativa.
^b. Empleando el error típico asintótico basado en la hipótesis nula.
^c. No se pueden calcular los estadísticos Kappa. Requieren una tabla simétrica de 2 vías en la que los valores de la primera variable sean idénticos a los valores de la segunda.

Figura 12-8

Pruebas de chi-cuadrado						
	Valor	gl		Sig. asintótica (bilateral)	Sig. exacta (bilateral)	
Chi-cuadrado de Pearson	79,277 ^a	2		,000		
Razón de verosimilitud	95,463	2		,000		
Asociación lineal por lineal	67,463	1		,000		
Prueba de McNemar						
N de casos válidos	474					^b

^a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 12,30.
^b. Sólo se efectuará el cálculo para tablas de PxP, donde P debe ser mayor que 1.

Figura 12-9

Estimación de riesgo	
	Valor
Razón de las ventajas para Sexo (Hombre / Mujer)	^a
a. No se puede calcular el estadístico Estimación de riesgo. Sólo se calcula para tablas 2x2 sin casillas vacías.	

Figura 12-10

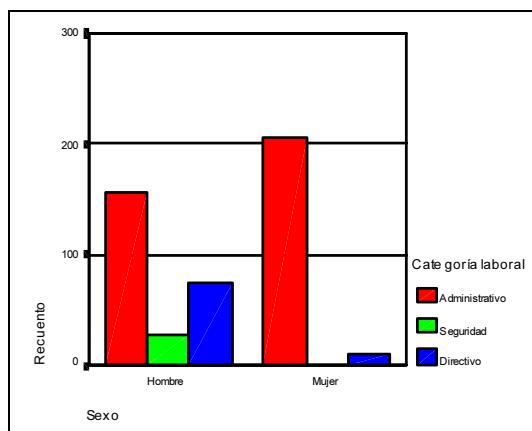


Figura 12-11

EL PROCEDIMIENTO RESUMIR

El procedimiento Resumir calcula estadísticos de subgrupo para las variables dentro de las categorías de una o más variables de agrupación. Se cruzan todos los niveles de las variables de agrupación. Puede elegir el orden en el que se mostrarán los estadísticos. También se muestran estadísticos de resumen para cada variable a través de todas las categorías. Los valores de los datos en cada categoría pueden mostrarse en una lista o suprimirse. Con grandes conjuntos de datos, tiene la opción de listar sólo los primeros n casos.

En cuanto a estadísticos, se obtiene la suma, número de casos, media, mediana, mediana agrupada, error típico de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación típica, varianza, curtosis, error típico de curtosis, asimetría, error típico de asimetría, porcentaje de la suma total, porcentaje del N total, porcentaje de la suma en, porcentaje de N en, media geométrica y media armónica.

En cuanto a los datos, las variables de agrupación son variables categóricas cuyos valores pueden ser numéricos o de cadena corta. El número de categorías debe ser razonablemente pequeño. Las otras variables deben poder ordenarse mediante rangos. Algunos de los estadísticos opcionales de subgrupo, como la media y la desviación típica, se basan en la teoría normal y son adecuados para variables cuantitativas con distribuciones simétricas. Los estadísticos robustos, tales como la mediana y el rango, son adecuados para las variables cuantitativas que pueden o no cumplir el supuesto de normalidad.

Para obtener resúmenes de casos, elija en los menús *Analizar* → *Informes* → *Resúmenes de casos* y seleccione las variables en la Figura 12-12. Seleccione una o más variables en la Figura 12-13. Como ejemplo, a partir del fichero MUNDO vamos a clasificar la población mundial (*poblac*), el índice de alfabetización (*alfabet*) y la mortalidad infantil (*mortalid*) por religiones (*relig*). Si lo desea, tiene la posibilidad de seleccionar una o más variables de agrupación para dividir los datos en subgrupos, pulsar en *Opciones* (Figura 12-14) para cambiar el título de los resultados o añadir un texto al pie debajo de los resultados o excluir los casos con valores perdidos, pulsar en *Estadísticos* (Figura 12-15) para acceder a estadísticos adicionales, seleccionar *Mostrar casos* para listar los casos en cada subgrupo. Por defecto, el sistema enumera sólo los 100 primeros casos del archivo. Puede aumentar o disminuir el valor de *Limitar los casos a los primeros*, o desactivar ese elemento para enumerar todos los casos. Al pulsar *Aceptar* en la Figura 12-13 se obtiene la salida del procedimiento (Figuras 12-16 y 12-17).

Analizar Gráficos Utilidades Ventana ?

Analizar	Gráficos	Utilidades	Ventana	?
Informes				
Resúmenes de casos...				
Informe de estadísticos en filas...				
Informe de estadísticos en columnas...				
54 Musulma. 75				
18 Musulma. 44				
85 Protest. 79				
77 Musulma. 70				
86 Católica 75				
68 Ortodoxa 75				
85 Protest. 80				
58 Católica 79				
83 Musulma. 74				

Figura 12-12



Figura 12-13

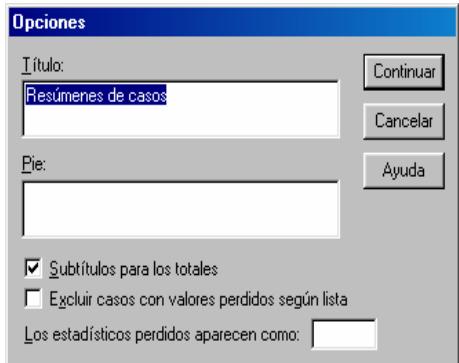


Figura 12-14



Figura 12-15

```

GET
FILE='C:\Archivos de programa\SPSS\MUNDO.sav'.
SUMMARIZE
/TABLES=alfabet mortalinf poblac BY relig
/FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=100
/TITLE='Resúmenes de casos'
/MISSING=VARIABLE
/CELLS=COUNT MEAN STDDEV KURT NPCT .

```

Resumir**Resumen del procesamiento de los casos***

	Casos					
	Incluidos		Excluidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Alfabetización (%) *						
Religión mayoritaria	97	97,0%	3	3,0%	100	100,0%
Mortalidad infantil (muertes por 1000 nacimientos vivos) *	99	99,0%	1	1,0%	100	100,0%
Religión mayoritaria						
Población x1000 *	99	99,0%	1	1,0%	100	100,0%
Religión mayoritaria						

a. Limitado a los primeros 100 casos.

Figura 12-16**Resúmenes de casos***

Religión	Animista	1	Alfabetización (%)	Mortalidad infantil (muertes por 1000 nacimientos vivos)		Población x1000
				Total	N	
Mayoritaria		2		64	77,0	13100
		3		40	113,0	2900
	Total		N	3	3	3
			Media	37,33	102,667	8666,67
			Desv. típ.	18,15	22,368	5220,09
			Curtosis	,	,	,
			% del total de N	,	,	,
Budista		1		36	112,0	10000
		2		99	27,7	23100
		3		77	5,8	5800
		4		99	4,4	125500
		5		93	37,0	59400
		6		91	5,1	20944
	Total		N	6	6	6
			Media	82,33	32,000	40790,67
			Desv. típ.	24,55	41,510	45609,74
			Curtosis	3,759	3,779	2,527
			% del total de N	,	,	,
Católica		1		95	26,6	33900
		2		99	6,7	8000
		3		99	7,7	10100

Figura 12-17

SPSS Y LOS MODELOS LOGARÍTMICO LINEALES

Selección del modelo

SPSS aborda el tratamiento de los modelos logarítmico lineales de un modo bastante ordenado. SPSS habilita tres procedimientos distintos cuya finalidad es seleccionar inicialmente un modelo para los datos (procedimiento *Selección del modelo*), poner a prueba cualquier modelo y analizarlo a fondo calculando sus parámetros (procedimiento *Loglineal general*) y distinguir entre variables dependientes e independientes (procedimiento *Análisis logit lineal*).

Evidentemente el procedimiento *Selección del modelo* es el primero que debe utilizarse, ya que tiene como finalidad realizar una aproximación exploratoria que nos ayude a identificar los términos necesarios en un modelo logarítmico lineal. Posteriormente se utilizará esta definición inicial del modelo para realizar las estimaciones definitivas.

En los ejercicios que acompañan este capítulo se desarrollará un caso completo para cada uno de los tres procedimientos citados.

Análisis loglineal general

Ya se hemos citado en el apartado anterior que el procedimiento *Análisis loglineal general* tiene como finalidad poner a prueba cualquier modelo identificado previamente mediante el procedimiento *Selección del modelo* y analizarlo a fondo estimando sus parámetros. Este procedimiento permite estimar modelos jerárquicos (saturados o no saturados) y modelos no jerárquicos.

Análisis logit

El procedimiento *Análisis logit* permite ajustar modelos logarítmico lineales distinguiendo entre variables dependientes e independientes. En este caso, una vez realizada la estimación, la combinación lineal de parámetros que se obtienen expresa los logaritmos de las razones esperadas de la variable dependiente (*odds*) y no la frecuencia esperada de la casilla.

Las variables dependientes e independientes son categóricas pudiendo admitirse también variables covariadas continuas (cada valor covariado representa la media de todos los casos pertenecientes a la casilla). Para ajustar este tipo de modelos logarítmico lineales se emplea el método de estimación iterativa de Newton Rhompson.

Ejercicio 12-1. Consideremos los datos procedentes de una encuesta sobre el uso del cinturón de seguridad en la que se miden las variables V1 = Momento de la conducción (con valores 1 = Día y 2 = Noche), V2 = Lugar de la conducción (con valores 3 = Autopista y 2 = Ciudad), V3 = Género del conductor (con valores 0 = Hombre y 1 = Mujer) y V4 = Uso del cinturón (con valores 0 = No usa y 1 = Sí usa). Los datos recogidos se tabulan según se indica en el problema (P es la frecuencia). Con esta información se trata de identificar qué factores se relacionan con el mayor o menor uso del cinturón de seguridad y de ver qué diferencias hay entre hombres y mujeres al usar el cinturón. También intenta analizarse qué diferencias hay entre el uso de cinturón por autopista o por ciudad o dependiendo de si es de día o de noche.

Estamos ante el clásico problema de aplicación de modelos logarítmico lineales, ya que disponemos de una tabla de contingencia de varias dimensiones con variables cualitativas.

Evidentemente el primer paso a dar es seleccionar un modelo que ajuste adecuadamente nuestros datos. La tabla de contingencia con los datos para la realización de este problema se presenta continuación:

V1	V2	V3	V4	P
Día	Autopista	Hombre	No usa	128,00
			Sí usa	589,00
		Mujer	No usa	29,00
			Sí usa	236,00
	Ciudad	Hombre	No usa	419,00
			Sí usa	312,00
		Mujer	No usa	132,00
			Sí usa	136,00
Noche	Autopista	Hombre	No usa	279,00
			Sí usa	499,00
		Mujer	No usa	44,00
			Sí usa	179,00
	Ciudad	Hombre	No usa	553,00
			Sí usa	249,00
		Mujer	No usa	98,00
			Sí usa	99,00

Se comienza introduciendo los datos en el editor de datos de SPSS como se indica en la figura 12-18. Como los datos vienen dados en forma de tabla de frecuencias, será necesario ponderar las variables mediante la columna de las frecuencias a través de *Datos → Ponderar casos* (figura 12-19) rellenando la pantalla *Ponderar casos* como se indica en la figura 12-20. A pulsar *Aceptar* ya tenemos los datos dispuestos para ser tratados mediante un modelo logarítmico lineal.

LOGLIN - Editor de datos SPSS

18 : p

	v1	v2	v3	v4	p
1	1,00	3,00	,00	,00	128,00
2	1,00	3,00	,00	1,00	589,00
3	1,00	3,00	1,00	,00	29,00
4	1,00	3,00	1,00	1,00	236,00
5	1,00	4,00	,00	,00	419,00
6	1,00	4,00	,00	1,00	312,00
7	1,00	4,00	1,00	,00	132,00
8	1,00	4,00	1,00	1,00	136,00
9	2,00	3,00	,00	,00	279,00
10	2,00	3,00	,00	1,00	499,00
11	2,00	3,00	1,00	,00	44,00
12	2,00	3,00	1,00	1,00	179,00
13	2,00	4,00	,00	,00	563,00
14	2,00	4,00	,00	1,00	249,00
15	2,00	4,00	1,00	,00	98,00
16	2,00	4,00	1,00	1,00	99,00

Figura 12-18

LOGLIN - Editor de datos SPSS

18 : p

	v1	p
1	1,00	128,00
2	1,00	589,00
3	1,00	29,00
4	1,00	236,00
5	1,00	419,00
6	1,00	312,00
7	1,00	132,00
8	1,00	136,00
9	2,00	279,00
10	2,00	499,00
11	2,00	44,00
12	2,00	1,00
13	2,00	179,00
14	2,00	563,00
15	2,00	249,00
16	2,00	98,00
	4,00	99,00

Figura 12-19

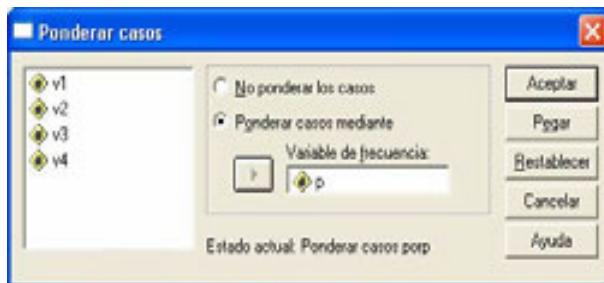


Figura 12-20

Para identificar los términos necesarios en el modelo logarítmico lineal utilizamos *Anализar → Loglineal → Selección de modelo* (figura 12-21) y rellenamos la pantalla de entrada como se indica en la figura 12-22. Las pantallas relativas a los botones *Modelo* y *Opciones* se llenan como se indica en las figuras 12-23 (para comenzar se elige el modelo saturado) y 12-24 (se piden frecuencias observadas y esperadas que serán coincidentes para el modelo saturado, residuos que serán nulos para el modelo saturado, estimaciones de los parámetros del modelo saturado con errores típicos e intervalos de confianza y la tabla de asociación que calcula las pruebas simultáneas y las pruebas de asociación parcial y que indicarán también la mayor o menor idoneidad de cada efecto para hacer una propuesta tentativa del modelo).

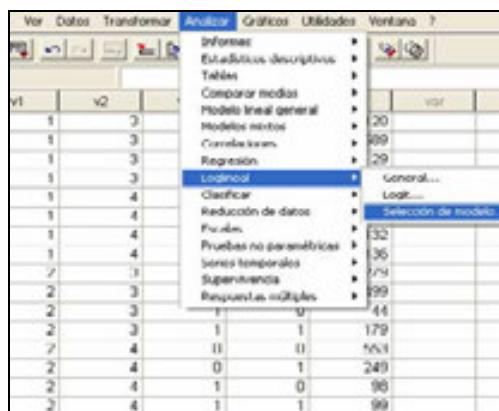


Figura 12-21



Figura 12-22

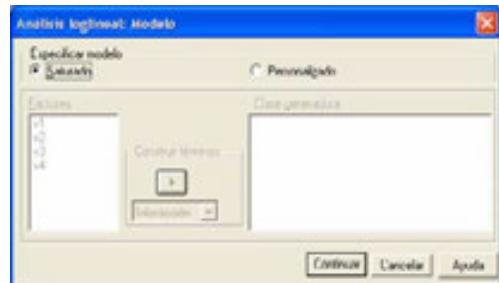


Figura 12-23

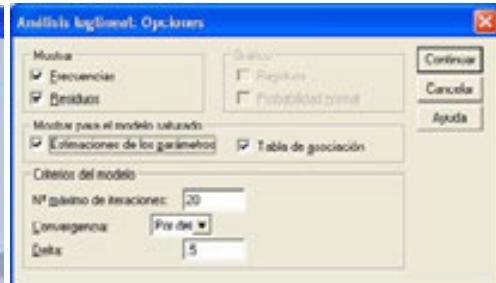


Figura 12-24

Al pulsar *Aceptar* en la figura 12-22 se obtiene la salida. En primer lugar se obtiene una descripción global del análisis indicando que el modelo es jerárquico, que no se han perdido observaciones maestrales manteniéndose los 3981 coches de la muestra y que la clase generadora (*design*) coincide con la iteración de orden 4 (la más alta).

```
* * * * * * * * * H I E R A R C H I C A L   L O G   L I N E A R * * * * * * *
```

DATA Information

```
16 unweighted cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
3981 weighted cases will be used in the analysis.
```

FACTOR Information

Factor	Level	Label
V1	2	
V2	2	
V3	2	
V4	2	

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

DESIGN 1 has generating class

V1*V2*V3*V4

Note: For saturated models ,500 has been added to all observed cells.
This value may be changed by using the CRITERIA = DELTA subcommand.

The Iterative Proportional Fit algorithm converged at iteration 1.
The maximum difference between observed and fitted marginal totals is ,000
and the convergence criterion is 346,921

A continuación se obtiene la tabla de frecuencias observadas, esperadas y residuales (iguales a cero para el modelo saturado). También se obtienen las pruebas de calidad del ajuste G^2 y χ^2 , cuyo valor es cero dado que el ajuste es perfecto en el modelo saturado.

Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
V1	Día				
V2	Autopist				
V3	Hombre				
V4	No usa	128,5	128,5	,00	,00
V4	Si usa	589,5	589,5	,00	,00
V3	Mujer				
V4	No usa	29,5	29,5	,00	,00
V4	Si usa	236,5	236,5	,00	,00
V2	Ciudad				
V3	Hombre				
V4	No usa	419,5	419,5	,00	,00
V4	Si usa	312,5	312,5	,00	,00
V3	Mujer				
V4	No usa	132,5	132,5	,00	,00
V4	Si usa	136,5	136,5	,00	,00
V1	Noche				
V2	Autopist				
V3	Hombre				
V4	No usa	279,5	279,5	,00	,00
V4	Si usa	499,5	499,5	,00	,00
V3	Mujer				
V4	No usa	44,5	44,5	,00	,00
V4	Si usa	179,5	179,5	,00	,00
V2	Ciudad				
V3	Hombre				
V4	No usa	553,5	553,5	,00	,00
V4	Si usa	249,5	249,5	,00	,00
V3	Mujer				
V4	No usa	98,5	98,5	,00	,00
V4	Si usa	99,5	99,5	,00	,00

Goodness-of-fit test statistics

Likelihood ratio chi square = ,00000 DF = 0 P = -INF
Pearson chi square = ,00000 DF = 0 P = -INF

A continuación se obtienen los resultados de las pruebas simultáneas en las que se observa que la interacción de cuatro variables no será idónea a la hora de proponer un modelo (p-valor mayor que 0,05). Si son idóneas las interacciones de tres variables, de dos y los efectos principales.

Tests that K-way and higher order effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
4	1	,880	,3482	,882	,3476	1
3	5	20,814	,0009	20,953	,0008	2
2	11	715,369	,0000	685,877	,0000	2
1	15	1948,302	,0000	1953,995	,0000	0

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	4	1232,933	,0000	1268,119	,0000	0
2	6	694,555	,0000	664,923	,0000	0
3	4	19,934	,0005	20,071	,0005	0
4	1	,880	,3482	,882	,3476	0

A continuación se obtienen los resultados de las pruebas de asociación parcial en las que se comprueba la significación de cada efecto individual, es decir, su contribución al ajuste del modelo. Mantendremos en el modelo los efectos con p-valor menor que 0,05 (sombreados en la tabla) siempre y cuando no entremos en contradicción con el modelo jerárquico, que siempre ha de contener todas las rams inferiores a una aceptada.

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

Tests of PARTIAL associations.

Effect Name	DF	Partial Chisq	Prob	Iter
V1*V2*V3	1	,000	1,0000	1
V1*V2*V4	1	21,952	,0000	1
V1*V3*V4	1	6,276	,0122	1
V2*V3*V4	1	6,921	,0085	1
V1*V2	1	12,124	,0005	2
V1*V3	1	11,272	,0008	2
V2*V3	1	2,684	,1014	2
V1*V4	1	73,442	,0000	2
V2*V4	1	555,497	,0000	2
V3*V4	1	51,906	,0000	2
V1	1	,091	,7624	2
V2	1	,057	,8120	2
V3	1	1136,773	,0000	1
V4	1	96,014	,0000	2

A continuación se obtienen las estimaciones de los parámetros bajo el modelo saturado. Como las variables son todas dicotómicas, para cada efecto se ofrece sólo un parámetro. Si una variable tuviese tres niveles se presentarían dos parámetros para el efecto principal de esa variable.

Estimates for Parameters.

V1*V2*V3*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	,0141771452	,02233	,63476	-,02960	,05795

V1*V2*V3

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	,0117915875	,02233	,52795	-,03198	,05557

V1*V2*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,0692267364	,02233	-3,09952	-,11300	-,02545

V1*V3*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,0461607182	,02233	-2,06678	-,08994	-,00238

V2*V3*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	,0150742427	,02233	,67493	-,02870	,05885

V1*V2

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,0817107867	,02233	-3,65848	-,12549	-,03793

V1*V3

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,0713016300	,02233	-3,19242	-,11508	-,02753

V2*V3

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	,0387214648	,02233	1,73370	-,00505	,08250

V1*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,1344748181	,02233	-6,02091	-,17825	-,09070

V2*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,4144758696	,02233	-18,55754	-,45825	-,37070

V3*V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	,1564623626	,02233	7,00537	,11269	,20024

V1

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,0116307373	,02233	-,52075	-,05541	,03215

V2

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,1059183582	,02233	-4,74234	-,14969	-,06214

V3

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	,6170048660	,02233	27,62547	,57323	,66078

V4

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-,2830485160	,02233	-12,67307	-,32682	-,23927

Por ejemplo, el primer parámetro de la interacción V1*V2*V3*V4 representa 1 celdilla que contiene los casos pertenecientes a la primera categoría de cada variable (*día / autopista / conductor varón / no lleva cinturón*). El segundo parámetro de esta interacción, que correspondería a la celdilla *día / autopista / conductor varón / sí lleva cinturón*, no se muestra, pero tendría igual valor con signo cambiado que el primer parámetro. Su valor Z absoluto tiene que ser superior o igual a 1,96 o a 2,57 para ser considerado significativamente distinto de cero al nivel 0,05 o 0,01 respectivamente, cosa que no ocurre, ya que el efecto interacción de las cuatro variables no era significativo. Tampoco lo es la interacción de tres variables V1*V2*V3.

La interacción V1*V3*V4 sí es significativa al 99% con valor absoluto de Z superior a 2,57 y cuyo intervalo de confianza no incluye al cero, cosa que sí ocurre en las iteraciones no significativas. Esta iteración corresponde a la casilla *día / autopista / no lleva cinturón*. Al tener signo negativo indica una menor frecuencia que su complementaria *día / autopista / no lleva cinturón*. Por lo tanto, los conductores que circulan de día por la autopista tienden mayoritariamente a llevar el cinturón puesto. De igual forma se interpretan los parámetros de los efectos restantes del modelo.

Si en la figura 12-22 marcamos *Utilizar la eliminación hacia atrás* (figura 12-25), SPSS emplea un procedimiento paso a paso que parte del modelo saturado y va eliminando los efectos no significativos empezando por el de mayor orden. El proceso continúa hasta que no pueda eliminarse ningún efecto más sin perder poder predictivo.



Figura 12-25

La salida adicional sería la siguiente:

```
* * * * * H I E R A R C H I C A L   L O G   L I N E A R * * * * * * * * *
Backward Elimination (p = ,050) for DESIGN 1 with generating class

V1*V2*V3*V4

Likelihood ratio chi square =      ,00000     DF = 0   P = -INF
-----  
If Deleted Simple Effect is          DF   L.R. Chisq Change   Prob   Iter
V1*V2*V3*V4                         1           ,880    ,3482      1

Step 1

The best model has generating class

V1*V2*V3
V1*V2*V4
V1*V3*V4
V2*V3*V4

Likelihood ratio chi square =      ,88015     DF = 1   P = ,348
-----  
If Deleted Simple Effect is          DF   L.R. Chisq Change   Prob   Iter
V1*V2*V3                           1           ,000  1,0000      1
Step 2

The best model has generating class

V1*V2*V4
V1*V3*V4
V2*V3*V4

Likelihood ratio chi square =      ,56220     DF = 2   P = ,755
-----  
If Deleted Simple Effect is          DF   L.R. Chisq Change   Prob   Iter
V1*V2*V4                           1           11,420  ,0007      2
V1*V3*V4                           1           6,918  ,0085      2
V2*V3*V4                           1           ,757  ,3843      1

Step 3

The best model has generating class

V1*V2*V4
V1*V3*V4
V2*V3

Likelihood ratio chi square =      1,31921    DF = 3   P = ,725
-----  
If Deleted Simple Effect is          DF   L.R. Chisq Change   Prob   Iter
V1*V2*V4                           1           13,844  ,0002      2
V1*V3*V4                           1           7,218  ,0072      2
V2*V3                             1           2,121  ,1453      1

Step 4

The best model has generating class
```

V1*V2*V4
V1*V3*V4

Likelihood ratio chi square = 3,44040 DF = 4 P = ,487

If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
V1*V2*V4	1		13,681	,0002		2
V1*V3*V4	1		7,743	,0054		2

Step 5

The best model has generating class

V1*V2*V4
V1*V3*V4

Likelihood ratio chi square = 3,44040 DF = 4 P = ,487

* * * * * H I E R A R C H I C A L L O G L I N E A R * * * * *

The final model has generating class

V1*V2*V4
V1*V3*V4

The Iterative Proportional Fit algorithm converged at iteration 0.
The maximum difference between observed and fitted marginal totals is ,000
and the convergence criterion is 346,921

El procedimiento ha llegado en 5 pasos a la conclusión de que el modelo final se genera por las clases V1*V2*V4 y V1*V3*V4. A continuación, para el modelo final el procedimiento muestra la tabla de frecuencias observadas, esperadas y residuales así como las pruebas de ajuste G^2 y χ^2 . Ambas pruebas indican que el modelo [V1V2V4, V1V3V4] es un buen modelo para ajustar los datos porque no presenta residuos significativos. Or otra parte, ningún residuo típico tiene un valor superior a 1,96 en valor absoluto, por lo que no hay casillas extremas.

Según el ajuste realizado es razonable proponer un modelo que incluya dos interacciones de tres factores, la interacción *Momento del día / Lugar de conducción / Uso del cinturón* y la interacción *Momento del día / Género del conductor / Uso del cinturón*. Los parámetros estimados para este modelo no se pueden obtener con este procedimiento, en cuyo caso hemos de recurrir al procedimiento *Análisis loglineal general*.

Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
V1	Día				
V2	Autopist				
V3	Hombre				
V4	No usa	128,0	128,0	,00	,00
V4	Si usa	589,0	589,0	,00	,00
V3	Mujer				
V4	No usa	29,0	29,0	,00	,00
V4	Si usa	236,0	236,0	,00	,00
V2	Ciudad				

V3	Hombre				
V4	No usa	419,0	419,0	,00	,00
V4	Si usa	312,0	312,0	,00	,00
V3	Mujer				
V4	No usa	132,0	132,0	,00	,00
V4	Si usa	136,0	136,0	,00	,00
V1	Noche				
V2	Autopist				
V3	Hombre				
V4	No usa	279,0	279,0	,00	,00
V4	Si usa	499,0	499,0	,00	,00
V3	Mujer				
V4	No usa	44,0	44,0	,00	,00
V4	Si usa	179,0	179,0	,00	,00
V2	Ciudad				
V3	Hombre				
V4	No usa	553,0	553,0	,00	,00
V4	Si usa	249,0	249,0	,00	,00
V3	Mujer				
V4	No usa	98,0	98,0	,00	,00
V4	Si usa	99,0	99,0	,00	,00

Goodness-of-fit test statistics

Likelihood ratio chi square = ,00000 DF = 4 P = 1,000
 Pearson chi square = ,00000 DF = 4 P = 1,000

Ejercicio 12-2. Con los datos del ejercicio anterior se trata de estimar el modelo logarítmico lineal identificado como adecuado para la información disponible. La estimación deberá conducir al estudio de qué factores se relacionan con el mayor o menor uso del cinturón de seguridad y de ver qué diferencias hay entre hombres y mujeres al usar el cinturón. También se estimarán las diferencias que hay entre el uso de cinturón por autopista o por ciudad o dependiendo de si es de día o de noche. Por lo tanto, la finalidad del ejercicio es la misma del ejercicio anterior, pero ahora ya más concisa

Identificado previamente el modelo mediante el procedimiento *Selección del modelo*, el procedimiento *Análisis loglineal general* permite poner a prueba cualquier modelo identificado previamente y analizarlo a fondo estimando sus parámetros. Este procedimiento permite estimar modelos jerárquicos (saturados o no saturados) y modelos no jerárquicos.

Se comienza introduciendo los datos en el editor de datos de SPSS como se indica en la figura 12-18. Como los datos vienen dados en forma de tabla de frecuencias, será necesario ponderar las variables mediante la columna de las frecuencias a través de *Datos → Ponderar casos* (figura 12-19) rellenando la pantalla *Ponderar casos* como se indica en la figura 12-20. A pulsar *Aceptar* ya tenemos los datos dispuestos para ser tratados mediante un modelo logarítmico lineal.

Para estimar el modelo logarítmico lineal utilizamos *Analizar → Loglineal → General* y rellenamos la pantalla de entrada como se indica en la figura 12-25.

Las pantallas relativas a los botones *Modelo* y *Opciones* se rellenan como se indica en las figuras 12-26 (se elige el modelo *Personalizado* y se pasan las variables a incluir en el modelo desde *Factores y Covariables* a *Términos del modelo*, teniendo presente que para formar cualquier interacción es necesario seleccionar a la vez todos sus términos en *Factores y covariables* y hacer doble clic con la selección activa) y 12-27 (se piden frecuencias, residuos, estimaciones, matriz de diseño y residuos corregidos con su probabilidad normal asociada). La figura 12-26 contiene todas las variables e interacciones que forman el modelo a estimar detectado como ideal con el procedimiento *Seleccionar modelo* que ya fue explicado anteriormente.



Figura 12-25



Figura 12-26

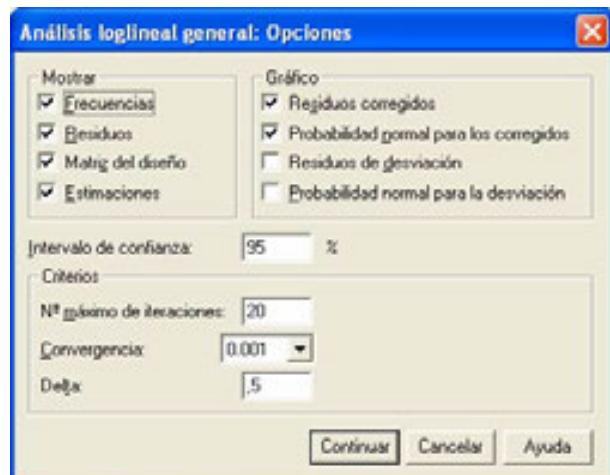


Figura 12-27

Al hacer clic en *Aceptar* en la figura 12-25 se obtiene la salida, cuya primera parte informa de las características de los datos y de las variables indicando los factores seleccionados y sus categorías.

GENERAL LOGLINEAR ANALYSIS

Data Information

16 cases are accepted.
0 cases are rejected because of missing data.
3981 weighted cases will be used in the analysis.
16 cells are defined.
0 structural zeros are imposed by design.
0 sampling zeros are encountered.

Variable Information

Factor	Levels	Value
V1	2	1 Día 2 Noche
V2	2	3 Autopista 4 Ciudad
V3	2	0 Hombre 1 Mujer
V4	2	0 No usa 1 Sí usa

Model and Design Information

Model: Poisson
Design: Constant + V1 + V2 + V3 + V4 + V1*V2 + V1*V3 + V1*V4 + V2*V4 + V3*V4
+ V1*V2*V4 + V1*V3*V4

La salida continua y numera todos los parámetros posibles señalando su correspondencia con los términos del diseño. Los términos redundantes (se pueden obtener a partir de otros) se señalan con una “x” y sus parámetros se igualarán a cero. De la misma forma se señalan para cada efecto los términos cuyos parámetros van a ser calculados y los que se igualarán a cero.

Posteriormente se presenta la matriz de diseño del modelo que nos indica si el parámetro está o no en la casilla de referencia (mediante un 1 o un 0 respectivamente). Por ejemplo, en la casilla *día / ciudad / hombre / no usa cinturón* están señalados con un 1 los parámetros 1, 2, 6, 8, 14, 18, 26 y 38. Si se sustituyen esos parámetros en la ecuación del modelo por los valores que aparecen en la tabla posterior de *Parámetros estimados*, se obtiene el logaritmo de la frecuencia esperada de dicha casilla. Elevando el número “e” a este valor se obtiene la frecuencia esperada para la casilla. De la misma forma se calculan las frecuencias esperadas para el resto de las casillas.

GENERAL LOGLINEAR ANALYSIS

Correspondence Between Parameters and Terms of the Design

Parameter	Aliased	Term
1		Constant
2		[V1 = 1]
3	x	[V1 = 2]
4		[V2 = 3]
5	x	[V2 = 4]
6		[V3 = 0]
7	x	[V3 = 1]
8		[V4 = 0]
9	x	[V4 = 1]
10		[V1 = 1]*[V2 = 3]
11	x	[V1 = 1]*[V2 = 4]
12	x	[V1 = 2]*[V2 = 3]
13	x	[V1 = 2]*[V2 = 4]
14		[V1 = 1]*[V3 = 0]
15	x	[V1 = 1]*[V3 = 1]
16	x	[V1 = 2]*[V3 = 0]
17	x	[V1 = 2]*[V3 = 1]
18		[V1 = 1]*[V4 = 0]
19	x	[V1 = 1]*[V4 = 1]
20	x	[V1 = 2]*[V4 = 0]
21	x	[V1 = 2]*[V4 = 1]
22		[V2 = 3]*[V4 = 0]
23	x	[V2 = 3]*[V4 = 1]
24	x	[V2 = 4]*[V4 = 0]
25	x	[V2 = 4]*[V4 = 1]
26		[V3 = 0]*[V4 = 0]
27	x	[V3 = 0]*[V4 = 1]
28	x	[V3 = 1]*[V4 = 0]
29	x	[V3 = 1]*[V4 = 1]
30		[V1 = 1]*[V2 = 3]*[V4 = 0]
31	x	[V1 = 1]*[V2 = 3]*[V4 = 1]
32	x	[V1 = 1]*[V2 = 4]*[V4 = 0]
33	x	[V1 = 1]*[V2 = 4]*[V4 = 1]
34	x	[V1 = 2]*[V2 = 3]*[V4 = 0]
35	x	[V1 = 2]*[V2 = 3]*[V4 = 1]
36	x	[V1 = 2]*[V2 = 4]*[V4 = 0]
37	x	[V1 = 2]*[V2 = 4]*[V4 = 1]
38		[V1 = 1]*[V3 = 0]*[V4 = 0]
39	x	[V1 = 1]*[V3 = 0]*[V4 = 1]
40	x	[V1 = 1]*[V3 = 1]*[V4 = 0]
41	x	[V1 = 1]*[V3 = 1]*[V4 = 1]
42	x	[V1 = 2]*[V3 = 0]*[V4 = 0]
43	x	[V1 = 2]*[V3 = 0]*[V4 = 1]
44	x	[V1 = 2]*[V3 = 1]*[V4 = 0]
45	x	[V1 = 2]*[V3 = 1]*[V4 = 1]

Note: 'x' indicates an aliased (or a redundant) parameter.
 These parameters are set to zero.

GENERAL LOGLINEAR ANALYSIS

Design Matrix

Factor	Value	Cell Structure	Parameter							
			1	2	4	6	8	10	14	18
V1	Día									
V2	Autopista									
V3	Hombre									
V4	No usa	1,000	1	1	1	1	1	1	1	1
V4	Sí usa	1,000	1	1	1	1	0	1	1	0
V3	Mujer									
V4	No usa	1,000	1	1	1	0	1	1	0	1
V4	Sí usa	1,000	1	1	1	0	0	1	0	0

V2		Ciudad									
V3		Hombre									
V4	No usa	1,000	1	1	0	1	1	0	1	1	1
V4	Sí usa	1,000	1	1	0	1	0	0	1	0	
V3		Mujer									
V4	No usa	1,000	1	1	0	0	1	0	0	1	
V4	Sí usa	1,000	1	1	0	0	0	0	0	0	
V1		Noche									
V2		Autopista									
V3		Hombre									
V4	No usa	1,000	1	0	1	1	1	0	0	0	0
V4	Sí usa	1,000	1	0	1	1	0	0	0	0	
V3		Mujer									
V4	No usa	1,000	1	0	1	0	1	0	0	0	
V4	Sí usa	1,000	1	0	1	0	0	0	0	0	
V2		Ciudad									
V3		Hombre									
V4	No usa	1,000	1	0	0	1	1	0	0	0	
V4	Sí usa	1,000	1	0	0	1	0	0	0	0	
V3		Mujer									
V4	No usa	1,000	1	0	0	0	1	0	0	0	
V4	Sí usa	1,000	1	0	0	0	0	0	0	0	
V1		Día									
V2		Autopista									
V3		Hombre									
V4	No usa	1,000	1	1	1	1					
V4	Sí usa	1,000	0	0	0	0					
V3		Mujer									
V4	No usa	1,000	1	0	1	0					
V4	Sí usa	1,000	0	0	0	0					
V2		Ciudad									
V3		Hombre									
V4	No usa	1,000	0	1	0	1					
V4	Sí usa	1,000	0	0	0	0					
V3		Mujer									
V4	No usa	1,000	0	0	0	0					
V4	Sí usa	1,000	0	0	0	0					
V1		Noche									
V2		Autopista									
V3		Hombre									
V4	No usa	1,000	1	1	0	0					
V4	Sí usa	1,000	0	0	0	0					
V3		Mujer									
V4	No usa	1,000	1	0	0	0					
V4	Sí usa	1,000	0	0	0	0					
V2		Ciudad									
V3		Hombre									
V4	No usa	1,000	0	1	0	0					
V4	Sí usa	1,000	0	0	0	0					
V3		Mujer									
V4	No usa	1,000	0	0	0	0					
V4	Sí usa	1,000	0	0	0	0					

Convergence Information

Maximum number of iterations: 20
 Relative difference tolerance: ,001
 Final relative difference: 2,45257E-05

Maximum likelihood estimation converged at iteration 3.

A continuación se presenta la tabla de frecuencias observadas y esperadas con porcentajes observados y esperados entre paréntesis. También se obtiene los residuos simples estandarizados y ajustados (ninguno significativo) y las pruebas G^2 y χ^2 .

GENERAL LOGLINEAR ANALYSIS

Table Information

Factor	Value	Observed Count	Expected Count	%
V1 Día				
V2 Autopista				
V3	Hombre			
V4	No usa	128,00 (3,22)	121,30 (3,05)	
V4	Sí usa	589,00 (14,80)	583,92 (14,67)	
V3	Mujer			
V4	No usa	29,00 (,73)	35,70 (,90)	
V4	Sí usa	236,00 (5,93)	241,08 (6,06)	
V2 Ciudad				
V3	Hombre			
V4	No usa	419,00 (10,52)	425,70 (10,69)	
V4	Sí usa	312,00 (7,84)	317,08 (7,96)	
V3	Mujer			
V4	No usa	132,00 (3,32)	125,30 (3,15)	
V4	Sí usa	136,00 (3,42)	130,92 (3,29)	
V1 Noche				
V2 Autopista				
V3	Hombre			
V4	No usa	279,00 (7,01)	275,91 (6,93)	
V4	Sí usa	499,00 (12,53)	494,29 (12,42)	
V3	Mujer			
V4	No usa	44,00 (1,11)	47,09 (1,18)	
V4	Sí usa	179,00 (4,50)	183,71 (4,61)	
V2 Ciudad				
V3	Hombre			
V4	No usa	553,00 (13,89)	556,09 (13,97)	
V4	Sí usa	249,00 (6,25)	253,71 (6,37)	
V3	Mujer			
V4	No usa	98,00 (2,46)	94,91 (2,38)	
V4	Sí usa	99,00 (2,49)	94,29 (2,37)	
V1 Día				
V2 Autopista				
V3	Hombre			
V4	No usa	6,70	1,45	,60
V4	Sí usa	5,08	,66	,21
V3	Mujer			
V4	No usa	-6,70	-1,45	-1,16
V4	Sí usa	-5,08	-,66	-,33
V2 Ciudad				
V3	Hombre			
V4	No usa	-6,70	-1,45	-,33
V4	Sí usa	-5,08	-,66	-,29
V3	Mujer			
V4	No usa	6,70	1,45	,59
V4	Sí usa	5,08	,66	,44
V1 Noche				
V2 Autopista				
V3	Hombre			
V4	No usa	3,09	,60	,19
V4	Sí usa	4,71	,70	,21
V3	Mujer			
V4	No usa	-3,09	-,60	-,46
V4	Sí usa	-4,71	-,70	-,35
V2 Ciudad				
V3	Hombre			
V4	No usa	-3,09	-,60	-,13
V4	Sí usa	-4,71	-,70	-,30
V3	Mujer			
V4	No usa	3,09	,60	,32
V4	Sí usa	4,71	,70	,48

Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	3,4403	4	,4870
Pearson	3,3659	4	,4986

A continuación obtenemos las estimaciones de los parámetros del modelo, sus errores típicos y su transformación en puntuaciones Z con el intervalo de confianza correspondiente (para $Z > 1,96$ hay significatividad al 95% y para $Z > 2,57$ hay significatividad al 99%). Los parámetros significativos aparecen sombreados.

GENERAL LOGLINEAR ANALYSIS

Parameter Estimates

Parameter	Estimate	SE	Z-value	Asymptotic Lower	95% CI Upper
1	4,5464	,0741	61,33	4,40	4,69
2	,3282	,0981	3,34	,14	,52
3	,0000
4	,6669	,0659	10,11	,54	,80
5	,0000
6	,9898	,0702	14,09	,85	1,13
7	,0000
8	,0065	,1142	,06	-,22	,23
9	,0000
10	-,0564	,0883	-,64	-,23	,12
11	,0000
12	,0000
13	,0000
14	-,1052	,0934	-1,13	-,29	,08
15	,0000
16	,0000
17	,0000
18	-,0504	,1543	-,33	-,35	,25
19	,0000
20	,0000
21	,0000
22	-1,3678	,0948	-14,43	-1,55	-1,18
23	,0000
24	,0000
25	,0000
26	,7782	,1148	6,78	,55	1,00
27	,0000
28	,0000
29	,0000
30	-,4983	,1436	-3,47	-,78	-,22
31	,0000
32	,0000
33	,0000
34	,0000
35	,0000
36	,0000
37	,0000
38	-,4398	,1582	-2,78	-,75	-,13
39	,0000
40	,0000
41	,0000
42	,0000
43	,0000
44	,0000
45	,0000

Los parámetros no significativos resultan ser los números 8, 10, 14 y 18 cuyos intervalos de confianza contienen el cero. Los valores absolutos de los parámetros indican la intensidad de la asociación. El efecto más intenso resulta ser el del parámetro 22 que corresponde a la casilla V2=3*V4=0 (*autopista / sin cinturón*). La subtabla del efecto V2*V4 con sus parámetros puede representarse como sigue:

		V2 Lugar de conducción	
		3 Autopista	4 Ciudad
V4 Uso cinturón	0 No	(p.22) -1,3678	(p.24) +1,3678
	1 Sí	(p.23) +1,3678	(p.25) -1,3678

El parámetro 22 con signo negativo indica que en autopista la tasa de conductores que no llevan cinturón es significativamente menor que las de los que sí la llevan. Sin embargo, en la ciudad es significativamente mayor la tasa de conductores sin cinturón.

El resto de los parámetros de efectos principales y de interacciones de dos y tres factores pueden analizarse de forma similar. Si observamos el parámetro 6 que señala el efecto principal V6, vemos que la tasa de conductores masculinos es superior a la tasa femenina. El parámetro 30 recoge la interacción V1*V2*V4 que indica que las menores tasas de conductores sin cinturón se dan circulando de día por la autopista. El parámetro 38 recoge la interacción V1*V3*V4 que muestra que la tasa de varones que no usan cinturón de día es menor que la de los que sí lo usan.

SPSS también produce salida gráfica en este procedimiento, obteniéndose el diagrama de dispersión entre residuos y frecuencias observadas o esperadas (figura 12-28) para comprobar si los residuos siguen algún patrón observable o son aleatorios, el gráfico de dispersión y los gráficos QQ de normalidad de los residuos (figuras 12-29 y 12-30).

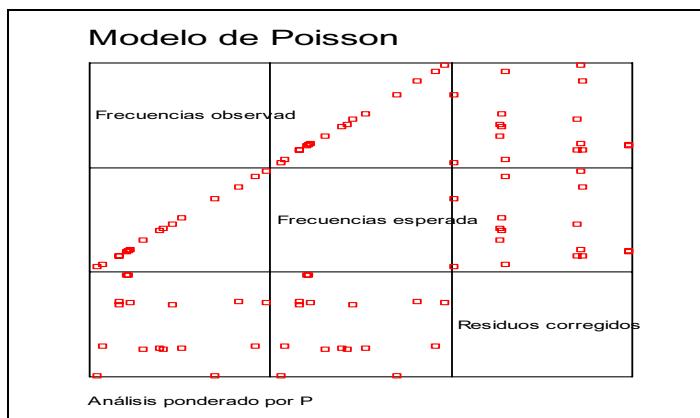


Figura 12-28

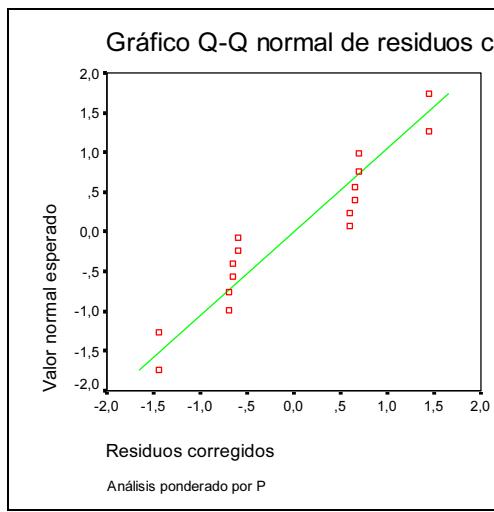


Figura 12-29

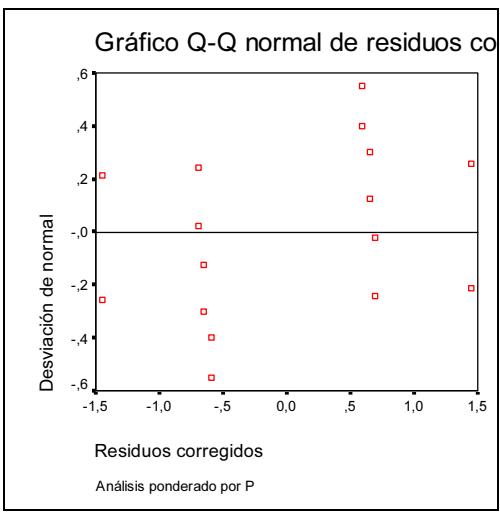


Figura 12-30

Ejercicio 12-3. Con los datos del ejercicio anterior se trata de estimar un modelo logit considerando el uso del cinturón V4 dependiendo del lugar de conducción V2 y del género del conductor V3..

El problema nos obliga esta vez a considerar el uso del cinturón V4 (variable dependiente o explicada) dependiendo del lugar de conducción V2 y del género del conductor V3 (variables independientes o explicativas). Para estimar el modelo logit utilizamos *Anализar* → *Loglineal* → *Logit* (figura 12-31) y rellenamos la pantalla de entrada como se indica en la figura 12-32. Las pantallas relativas a los botones *Modelo* y *Opciones* se llenan como se indica en las figuras 12-33 (se elige el modelo *Personalizado* y se incluyen sólo los efectos principales como términos del modelo) y 12-34 (se piden frecuencias, residuos, estimaciones, matriz de diseño y residuos corregidos con su probabilidad normal asociada).

Al hacer clic en *Aceptar* en la figura 12-32, se observa que la salida es muy similar a la habitual. El modelo incluye las interacciones de cada variable independiente con la dependiente (parte sustancial para la interpretación) y un efecto para la variable dependiente solicitado en la figura 12-33. El modelo incluye también una constante para cada combinación de niveles de las variables independientes. La matriz de diseño tiene el mismo sentido que en el procedimiento anterior. También aparecen los residuales (que deben ser significativamente distintos de cero para que el modelo se ajuste bien a los datos) y las pruebas de calidad del ajuste (los índices residuales G^2 y χ^2 deben de ser no significativos para un buen ajuste).

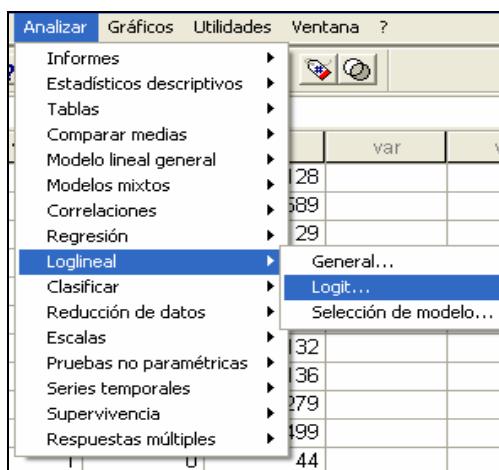


Figura 12-31

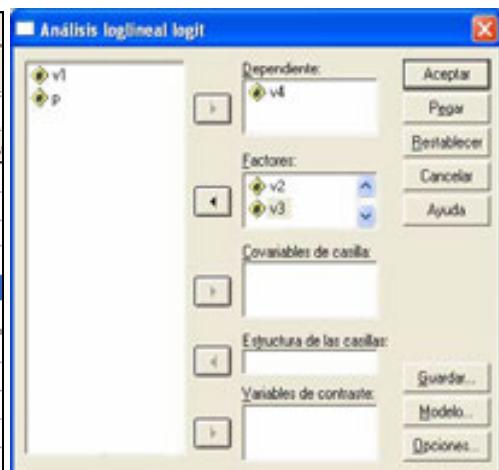


Figura 12-32

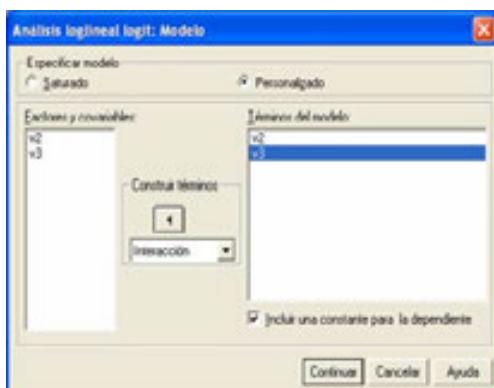


Figura 12-33



Figura 12-34

La salida es la siguiente:

```

-----  

GENERAL LOGLINEAR ANALYSIS  

-----  

Data Information  

16 cases are accepted.  

0 cases are rejected because of missing data.  

3981 weighted cases will be used in the analysis.  

8 cells are defined.  

0 structural zeros are imposed by design.  

0 sampling zeros are encountered.  

-----  

Variable Information

```

Factor	Levels	Value
V4	2	0 No usa 1 Si usa
V2	2	3 Autopista 4 Ciudad
V3	2	0 Hombre 1 Mujer

Model and Design Information

Model: Multinomial Logit

Design: Constant + V4 + V4*V2 + V4*V3

Note: There is a separate constant term for each combination of levels of the independent factors.

Correspondence Between Parameters and Terms of the Design

Parameter Aliased Term

1	Constant for [V2 = 3]*[V3 = 0]
2	Constant for [V2 = 3]*[V3 = 1]
3	Constant for [V2 = 4]*[V3 = 0]
4	Constant for [V2 = 4]*[V3 = 1]
5	[V4 = 0]
6	x [V4 = 1]
7	[V4 = 0]*[V2 = 3]
8	x [V4 = 0]*[V2 = 4]
9	x [V4 = 1]*[V2 = 3]
10	x [V4 = 1]*[V2 = 4]
11	[V4 = 0]*[V3 = 0]
12	x [V4 = 0]*[V3 = 1]
13	x [V4 = 1]*[V3 = 0]
14	x [V4 = 1]*[V3 = 1]

Note: 'x' indicates an aliased (or a redundant) parameter.
These parameters are set to zero.

Design Matrix

Factor	Value	Structure	Parameter					
			1	2	3	4	5	7
V2	Autopista							
V3	Hombre							
V4	No usa	1,000	1	0	0	0	1	1
V4	Si usa	1,000	1	0	0	0	0	0
V3	Mujer							
V4	No usa	1,000	0	1	0	0	1	1
V4	Si usa	1,000	0	1	0	0	0	0
V2	Ciudad							
V3	Hombre							
V4	No usa	1,000	0	0	1	0	1	0
V4	Si usa	1,000	0	0	1	0	0	0
V3	Mujer							
V4	No usa	1,000	0	0	0	1	1	0
V4	Si usa	1,000	0	0	0	1	0	0

Convergence Information

Maximum number of iterations: 20
 Relative difference tolerance: ,001
 Final relative difference: 4,52257E-05

Maximum likelihood estimation converged at iteration 3.

-

 GENERAL LOGLINEAR ANALYSIS

Table Information

Factor	Value	Observed	Expected		
		Count	%	Count	%
V2 Autopista					
V3	Hombre				
V4	No usa	407,00	(27,22)	400,97	(26,82)
V4	Si usa	1088,00	(72,78)	1094,03	(73,18)
V3	Mujer				
V4	No usa	73,00	(14,96)	79,03	(16,19)
V4	Si usa	415,00	(85,04)	408,97	(83,81)
V2 Ciudad					
V3	Hombre				
V4	No usa	972,00	(63,41)	978,03	(63,80)
V4	Si usa	561,00	(36,59)	554,97	(36,20)
V3	Mujer				
V4	No usa	230,00	(49,46)	223,97	(48,17)
V4	Si usa	235,00	(50,54)	241,03	(51,83)

Table Information

Factor	Value	Adj.	Dev.	
		Resid.	Resid.	Resid.
V2 Autopista				
V3	Hombre			
V4	No usa	6,03	1,04	3,49
V4	Si usa	-6,03	-1,04	-3,47
V3	Mujer			
V4	No usa	-6,03	-1,04	-3,40
V4	Si usa	6,03	1,04	3,49
V2 Ciudad				
V3	Hombre			
V4	No usa	-6,03	-1,04	-3,47
V4	Si usa	6,03	1,04	3,48
V3	Mujer			
V4	No usa	6,03	1,04	3,50
V4	Si usa	-6,03	-1,04	-3,45

 GENERAL LOGLINEAR ANALYSIS

Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	1,1003	1	,2942
Pearson	1,0893	1	,2966

Analysis of Dispersion

Source of Dispersion Entropy Concentration DF

Due to Model	300,5360	283,0592	2
Due to Residual	2410,8764	1659,6276	3978
Total	2711,4124	1942,6868	3980

Measures of Association

Entropy = ,1108
Concentration = ,1457

Parameter Estimates

Constant	Estimate
1	6,9976
2	6,0136
3	6,3189
4	5,4849

Note: Constants are not parameters under multinomial assumption.
Therefore, standard errors are not calculated.

Parameter	Estimate	SE	Z-value	Asymptotic 95% CI	
				Lower	Upper
5	-,0734	,0783	-,94	-,23	,08
6	,0000
7	-1,5704	,0703	-22,35	-1,71	-1,43
8	,0000
9	,0000
10	,0000
11	,6401	,0841	7,61	,48	,80
12	,0000
13	,0000
14	,0000

GENERAL LOGLINEAR ANALYSIS

Covariance Matrix of Parameter Estimates

Parameter	5	7	11
5	,0061		
7	-,0018	,0049	
11	-,0053	-4,44E-04	,0071

Aliased parameters and constants are not shown.

Correlation Matrix of Parameter Estimates

Parameter	5	7	11
5	1,0000		
7	-,3258	1,0000	
11	-,8083	-,0751	1,0000

Aliased parameters and constants are not shown.

Observaremos especialmente la interacción de cada variable independiente con la dependiente. El efecto sobre la dependiente debido al factor *Lugar* se ve en el parámetro 7 y el debido al factor *Género* en el parámetro 11. Ambos efectos son significativos, según muestran sus puntuaciones Z y sus intervalos de confianza.

Queda claro que llevar o no llevar cinturón depende significativamente del género del conductor y del lugar por donde se circule. La razón logarítmica de no llevar cinturón en autopista es -1,5704 veces la de llevarlo. Como $e^{-1,5704} = 0,21$, se puede decir también que la razón de no llevar cinturón en una autopista es 0,21 veces la de sí llevarlo. Respecto al género, la razón logarítmica de que los varones no lleven cinturón es 0,64 veces la de llevarlo. Como $e^{0,64} = 1,89$, la razón de que los varones no lleven cinturón es 1,89 veces la de llevarlo. La asociación entre las variables independientes y la dependiente se aproxima por el índice de entropía de Shannon y por el índice de concentración de Gini que aparecen en la tabla *Análisis de la dispersión*.

Ejercicio 12-4. Se pregunta a 50 economistas, 40 ingenieros y 10 abogados si creen que la bolsa en el próximo mes va a bajar, subir o permanecer igual. El 20% de los economistas opina que subirá, mientras que el 40% de ellos piensa que bajará. El 50% de los ingenieros se inclina por que permanecerá igual, y tan sólo el 5% cree que bajará. Por último, la mitad de los abogados se decanta por la subida y la otra mitad cree que bajará.

- Resumir los datos en la variable bidimensional que cruza la profesión con el pronóstico y presentar la tabla de frecuencias correspondiente y el gráfico de barras agrupadas para las variables unidimensionales.
- ¿Existe relación entre los pronósticos sobre la evolución del mercado bursátil y la profesión del encuestado?
- Hallar las distribuciones marginales del atributo profesión y del atributo pronóstico y realizar un diagrama de barras para cada atributo.
- Hallar la distribución del atributo pronóstico condicionada a la profesión de ingeniero y realizar un diagrama de sectores para la condicionada.

Comenzaremos calculando la tabla de contingencia relativa al problema, que quedaría como sigue:

Pronóstico →\nProfesión\n↓	Baja	Igual	Sube	Total
Economista	20	20	10	50
Ingeniero	2	20	18	40
Abogado	5	0	5	10
Total	27	40	33	100

La siguiente tarea será introducir los datos de los dos atributos como dos variables en el editor de SPSS. Denominamos P a la variable pronóstico y F a la variable profesión. P puede tomar los valores B (baja), I (igual) y S (sube). F puede tomar los valores E (economista), I (Ingeniero) y A (abogado). El valor EB de la variable bidimensional lo introducimos 20 veces (la E en la columna de la variable F y la B en la columna de la variable P), el valor EI 20 veces, el valor ES 10 veces, el valor IB 10 veces, el valor II 20 veces, el valor IS 18 veces, el valor AB 5 veces y el valor AS 5 veces.

Para estudiar la variable bidimensional y ver la relación entre las variables unidimensionales categóricas, rellenamos la pantalla de entrada del procedimiento *Tablas de contingencia* tal y como se indica en la Figura 12-35. Pulsamos en el botón *Estadísticos* y en la Figura 12-36 elegimos las medidas de asociación. Pulsamos en *Continuar* y *Aceptar* y obtenemos la salida del procedimiento en la que se observa su sintaxis y la tabla de contingencia P (distribución bidimensional), así como el gráfico de barras agrupadas para las variables (Figuras 12-37 y 12-38).

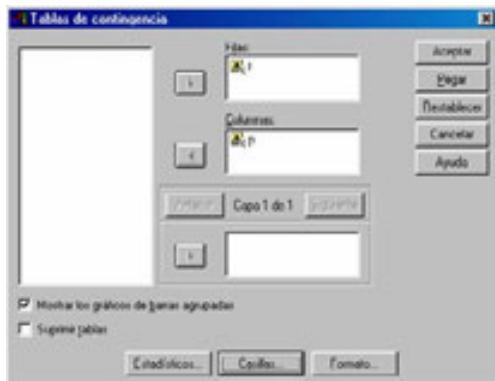


Figura 12-35

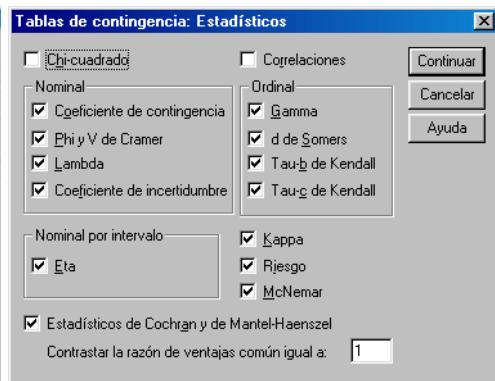


Figura 12-36

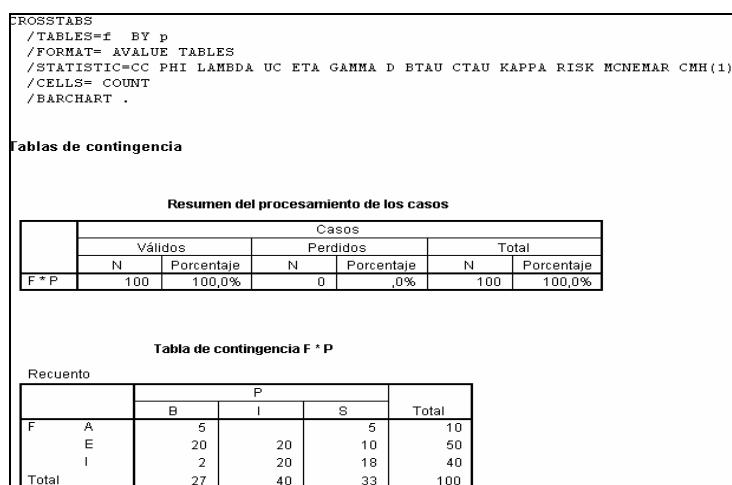


Figura 12-37

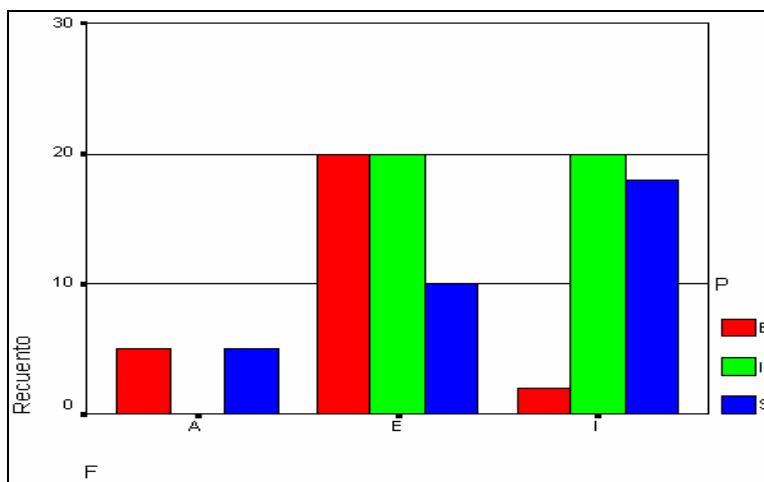


Figura 12-38

El resto de la salida del procedimiento mide el grado de asociación de P y F. Los valores obtenidos para las **medidas de asociación globales o simétricas** (Figura 12-39), como el coeficiente de contingencia de Pearson (0,428), la V de Cramer (0,335) la Phi (0,474), la Gamma (0,438) y las Taus (0,288 y 0,267) indican que el grado de asociación es muy pequeño, aunque evidentemente los p-valores de los contrastes constatan la existencia de una cierta asociación entre las variables F y P al 95% de coeficiente de confianza.

En la Figura 12-40, las **medidas de asociación direccionales**, como la Lambda, la Tau de Goodman y Kruskall, el coeficiente de incertidumbre, y la D de Somer, presentan el p-valor de los contrastes de asociación correspondientes. Si el p-valor es menor que 0,05, se acepta la hipótesis alternativa de existencia de asociación entre las dos variables. Para nuestro ejemplo se acepta la asociación en la mayoría de los casos.

Medidas simétricas					
		Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Phi	,474			,000
	V de Cramer	,335			,000
	Coeficiente de contingencia	,428			,000
Ordinal por ordinal	Tau-b de Kendall	,288	,093	3,120	,002
	Tau-c de Kendall	,267	,086	3,120	,002
	Gamma	,438	,137	3,120	,002
Medida de acuerdo N de casos válidos	Kappa	,0			

^a. Asumiendo la hipótesis alternativa.
^b. Empleando el error típico asintótico basado en la hipótesis nula.
^c. No se pueden calcular los estadísticos Kappa. Requieren una tabla simétrica de 2 vías en la que los valores de la primera variable sean idénticos a los valores de la segunda.

Figura 12-39

Medidas direccionales ^e				
		Valor	Error típ. asint. ^a	T aproximada ^b
Nominal por nominal	Lambda	Simétrica	,118	,072
		F dependiente	,160	,097
		P dependiente	,083	,107
	Tau de Goodman y Kruskal	F dependiente	,129	,042
		P dependiente	,107	,025
	Coeficiente de incertidumbre	Simétrica	,145	,037
		F dependiente	,156	,039
		P dependiente	,135	,035
Ordinal por ordinal	d de Somer	Simétrica	,288	,092
		F dependiente	,270	,086
		P dependiente	,307	,100

a. Asumiendo la hipótesis alternativa.
 b. Empleando el error típico asintótico basado en la hipótesis nula.
 c. Basado en la aproximación chi-cuadrado.
 d. Probabilidad del chi-cuadrado de la razón de verosimilitud.
 e. El estadístico ETA sólo es aplicable a datos numéricos.

Figura 12-40

Para hallar las distribuciones marginales de P y F, rellenamos la pantalla de entrada del procedimiento *Frecuencias* como se indica en la Figura 12-41. Pulsamos en el botón *Gráficos* y rellenamos la pantalla resultante como se indica en la Figura 12-42. Se pulsa *Continuar* y *Aceptar* y se obtienen las distribuciones de frecuencias de P y F y sus diagramas de barras (Figuras 12-43 a 12-45).

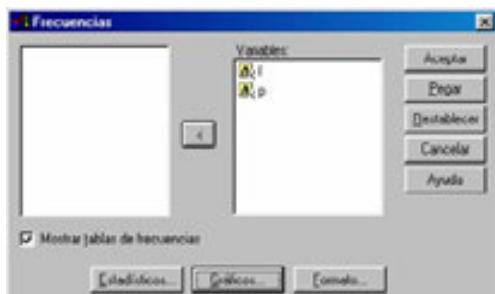


Figura 12-41



Figura 12-42

Tabla de frecuencia					
F					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	A	10	10,0	10,0	10,0
	E	50	50,0	50,0	60,0
	I	40	40,0	40,0	100,0
Total		100	100,0	100,0	

P					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	B	27	27,0	27,0	27,0
	I	40	40,0	40,0	67,0
	S	33	33,0	33,0	100,0
Total		100	100,0	100,0	

Figura 12-43

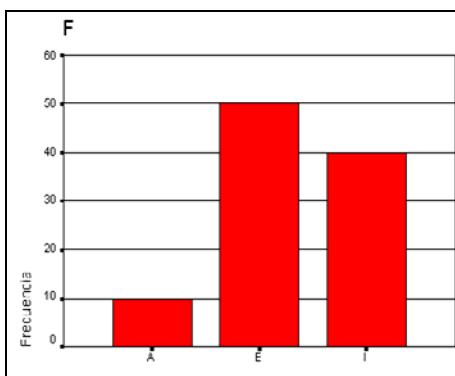


Figura 12-44

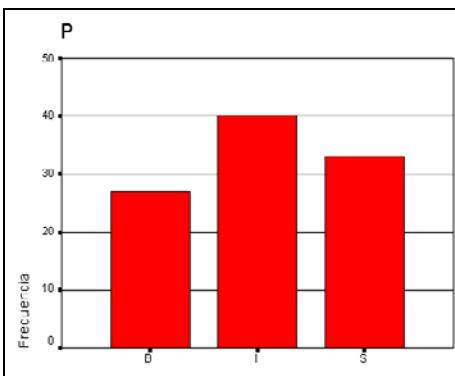


Figura 12-45

Para hallar la distribución de P condicionada a F rellenamos la pantalla de entrada del procedimiento *Resúmenes de casos* tal y como se indica en la Figura 12-46. Pulsamos en el botón *Estadísticos* se elige la media, la varianza y la mediana. A continuación pulsamos en *Aceptar* y obtenemos la salida del procedimiento en la que se observa la distribución condicionada de P a todos los valores de F (Figura 12-47).



Figura 12-46

Resúmenes de casos ^a		
F	A	P
		Total N
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
Total	1	10
	2	
	3	
	4	
	5	
	6	
	7	
	8	
	9	
	10	
	11	
	12	
	13	
	14	

Figura 12-47

CLASIFICACIÓN Y SEGMENTACIÓN MEDIANTE ANÁLISIS CLUSTER

CONCEPTO DE ANÁLISIS CLUSTER

El término *análisis cluster* se utiliza para definir una serie de técnicas, fundamentalmente algoritmos, que tienen por objeto la búsqueda de grupos similares de individuos o de variables que se van agrupando en conglomerados. Dada una muestra de individuos, de cada uno de los cuales se dispone de una serie de observaciones, el análisis cluster sirve para clasificarlos en grupos lo más homogéneos posible en base a las variables observadas. Los individuos que quedan clasificados en el mismo grupo serán tan similares como sea posible.

La palabra *cluster*, que define estas técnicas, se podría traducir por grupo, conglomerado, racimo, apiñarse, etc. El análisis cluster se usa en biología para clasificar animales y plantas, conociéndose con el nombre de *taxonomía numérica*. Otros nombres asignados al mismo concepto son análisis de *conglomerados*, *análisis tipológico*, *clasificación automática* y otros. Todos ellos pueden funcionar como sinónimos. En los paquetes estadísticos más habituales y en muchos trabajos en castellano suele aparecer el nombre de *cluster analysis*. Para Sokal y Sneath (1963), dos de los autores que más han influido en el desarrollo del análisis cluster, la clasificación es uno de los procesos fundamentales de la ciencia, ya que los fenómenos deben ser ordenados para que podamos entenderlos. Tanto el análisis cluster como el análisis discriminante sirven para clasificar individuos en categorías. La diferencia principal entre ellos estriba en que en el análisis discriminante se conoce a priori el grupo de pertenencia, mientras que el análisis cluster sirve para ir formando grupos homogéneos de conglomerados.

El análisis cluster es un método estadístico multivariante de clasificación automática de datos. A partir de una tabla de casos-variables, trata de situar los casos (individuos) en grupos homogéneos, conglomerados o clusters, no conocidos de antemano pero sugeridos

por la propia esencia de los datos, de manera que individuos que puedan ser considerados similares sean asignados a un mismo cluster, mientras que individuos diferentes (disimilares) se localicen en clusters distintos. La diferencia esencial con el análisis discriminante estriba en que en este último es necesario especificar previamente los grupos por un camino objetivo, ajeno a la medida de las variables en los casos de la muestra. El análisis cluster define grupos tan distintos como sea posible en función de los propios datos.

El enorme campo de aplicación en numerosas disciplinas, que se inició con la clasificación de las especies biológicas, ha propiciado la diversificación de este análisis, con denominaciones específicas tales como taxonomía numérica, taximetría, nosología, nosografía, morfometría, tipología, botriología, etc. La creación de grupos basados en similaridad de casos exige una definición de este concepto, o de su complementario «distancia» entre individuos. La variedad de formas de medir diferencias multivariadas o distancias entre casos proporciona diversas posibilidades de análisis. El empleo de ellas, y el de las que continuamente siguen apareciendo, así como de los algoritmos de clasificación, o diferentes reglas matemáticas para asignar los individuos a distintos grupos, depende del fenómeno estudiado y del conocimiento previo de posible agrupamiento que de él se tenga.

Puesto que la utilización del análisis cluster ya implica un desconocimiento o conocimiento incompleto de la clasificación de los datos, el investigador ha de ser consciente de la necesidad de emplear varios métodos, ninguno de ellos incuestionable, con el fin de contrastar los resultados.

Existen dos grandes tipos de análisis de clusters: Aquéllos que asignan los casos a grupos diferenciados que el propio análisis configura, sin que unos dependan de otros, se conocen como **no jerárquicos**, y aquéllos que configuran grupos con estructura arborescente, de forma que clusters de niveles más bajos van siendo englobados en otros de niveles superiores, se denominan **jerárquicos**. Los métodos no jerárquicos pueden, a su vez, producir **clusters disjuntos** (cada caso pertenece a un y sólo un cluster), o bien **solapados** (un caso puede pertenecer a más de un grupo). Estos últimos, de difícil interpretación, son poco utilizados.

Una vez finalizado un análisis de clusters, el investigador dispondrá de su colección de casos agrupada en subconjuntos jerárquicos o no jerárquicos. Podrá aplicar técnicas estadísticas comparativas convencionales siempre que lo permita la relevancia práctica de los grupos creados; así como otras pruebas multivariantes, para las que ya contará con una variable dependiente «grupo», aunque haya sido artificialmente creada.

De este modo, el horizonte de la investigación podría ampliarse, por ejemplo, con la aplicación de regresión logística y análisis discriminante con posibles nuevas variables independientes (utilizar las mismas que han servido para la confección de los grupos no sería una práctica correcta). También serían aplicables pruebas de asociación y análisis de correspondencias.

El análisis cluster se puede utilizar para agrupar individuos (casos) y también para agrupar variables. En lo que sigue, cuando nos refiramos a grupos de individuos (o casos), debe sobreentenderse que también nos referimos a conjuntos de variables. El proceso es idéntico tanto si se agrupan individuos como variables.

Antes de iniciar un análisis cluster deben tomarse tres decisiones: selección de las variables relevantes para identificar a los grupos, elección de la medida de proximidad entre los individuos y elección del criterio para agrupar individuos en conglomerados. Es decisiva la selección de las variables que realmente sean relevantes para identificar a los grupos, de acuerdo con el objetivo que se pretenda lograr en el estudio. De lo contrario, el análisis carecerá de sentido. Para la selección de la medida de proximidad es conveniente estar familiarizado con este tipo de medidas, básicamente similitudes y distancias, ya que los conglomerados que se forman lo hacen en base a las proximidades entre variables o individuos. Puesto que los grupos que se forman en cada paso depende de la proximidad, distintas medidas de proximidad pueden dar resultados distintos para los mismos datos. Para elegir el criterio de agrupación conviene conocer, como mínimo los principales métodos de análisis cluster.

DISTANCIAS Y SIMILITUDES

La proximidad expresa la semejanza que existe entre individuos o variables. Es decir, es el grado de asociación que existe entre ellos. Las proximidades pueden medir la distancia o la similitud (similaridad) entre individuos o variables. El valor que se obtiene en una medida de distancia es tanto mayor cuanto más alejados están los individuos o puntos entre los que se mide. En las similitudes, al contrario de las distancias, el valor que se obtiene es tanto mayor cuanto más próximos están los elementos considerados. La correlación de Pearson y los coeficientes de Spearman y de Kendall son índices de similitud.

Matemáticamente se da el nombre de distancia entre dos puntos A y B, a toda medida que verifique los axiomas siguientes:

1. $d(A,B) \geq 0$ y $d(A,A) = 0$
2. $d(A,B) = d(B,A)$
3. $d(A,B) \leq d(A,C) + d(C,B)$

Un primer ejemplo de distancia es la *distancia euclídea* que se define como:

$$d(A, B) = \sqrt{\sum_i (A_i - B_i)^2}$$

Un segundo ejemplo de distancia es la distancia D^2 de Mahalanobis. Originariamente se utilizó para calcular la distancia entre poblaciones. La D^2 es la distancia al cuadrado entre los centroides de dos poblaciones. Recordemos que el centroide de una población es el centro de gravedad de esta población en base a un conjunto de variables (vector de las medias de las variables). El centroide es en el análisis multivariable lo que la media es en el análisis univariable. Cronológicamente la D^2 de Mahalanobis está considerada como la primera técnica de análisis multidimensional. Su autor la formuló en 1927 y se divulgó algo más tarde (Mahalanobis, 1936). Las aplicaciones de esta técnica han ido modificando su carácter original, estableciéndose una relación entre teoría y práctica, que se ha ido enriqueciendo mutuamente. Las aplicaciones prácticas han descubierto nuevos aspectos, que generalizados en el plano formal, han dado lugar a nuevos procedimientos e interpretaciones.

Rao (1952) dio un gran impulso a esta medida con el desarrollo de proyecciones y representaciones gráficas de tipo geométrico. Las proyecciones permitieron posteriormente constatar y demostrar que la D^2 de Mahalanobis y el análisis discriminante constituyen dos aspectos del mismo proceso, en el sentido de dos formas de cálculo diferente para un mismo tipo de análisis. La D^2 de Mahalanobis permite situar poblaciones en un espacio de v dimensiones, siendo v el número de variables consideradas en el estudio. Entre las aplicaciones de la D^2 se encuentran los contrastes entre poblaciones, establecer la distancia entre dos individuos, calcular la distancia de un individuo al centroide de su grupo, etc.

Sean p poblaciones de n_1, n_2, \dots, n_p individuos cada una. En cada población se conocen v variables x_1, x_2, \dots, x_v , de modo que a cada población k le corresponde una matriz de observaciones de orden $n_k \times v$. Se dispone por tanto de p matrices de orden $n_k \times v$ con $k=1, \dots, p$. A partir de estos datos, y en notación matricial, Mahalanobis define la distancia entre los centroides de los grupos p y q (**distancia entre las poblaciones p y q**) como:

$$D_{pq}^2 = (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)$$

Los vectores $\boldsymbol{\mu}_p$ y $\boldsymbol{\mu}_q$ son vectores columna que contienen las medias de las variables de los grupos respectivos y $\boldsymbol{\Sigma}$ es la matriz de varianzas covarianzas intragrupos de los grupos conjuntamente.

A partir de la D^2 se puede estimar la F de Fisher y utilizarla como prueba de contraste para dos poblaciones:

$$F = D^2 \frac{n_p n_q (n_p + n_q - v - 1)}{(n_p + n_q)(n_p + n_q - 2)v} \rightarrow F_{v, n_p + n_q - v - 1}$$

Teniendo en cuenta que normalmente se trabaja con muestras, la fórmula que se utiliza para la D^2 es:

$$D_{pq}^2 = (\bar{\mathbf{X}}_p - \bar{\mathbf{X}}_q)' \mathbf{S}^{-1} (\bar{\mathbf{X}}_p - \bar{\mathbf{X}}_q)$$

donde las medias poblacionales se han estimado por medias muestrales y la varianza poblacional se ha estimado por la cuasivarianza muestral S^2 .

La D^2 de Mahalonobis puede utilizarse también para medir la *distancia entre dos individuos A y B* de la forma siguiente:

$$D_{AB}^2 = (\mathbf{X}_A - \mathbf{X}_B)' \Sigma^{-1} (\mathbf{X}_A - \mathbf{X}_B)$$

La D^2 de Mahalonobis también puede utilizarse para medir la *distancia de un individuo A al centroide de su grupo* de la forma siguiente:

$$D^2 = (\mathbf{X}_A - \bar{\mathbf{X}})' \Sigma^{-1} (\mathbf{X}_A - \bar{\mathbf{X}})$$

Los individuos con mayor D^2 son los más lejanos del centro de su grupo, con lo que esta distancia puede utilizarse para detectar individuos con puntuaciones extremas (*outliers*).

Un ejemplo de distancia entre dos variables x e y es la *distancia de Manhattan o City-block* que se define como:

$$B(x, y) = \sum_i |x_i - y_i|$$

Otro ejemplo de distancia entre dos variables x e y es la *distancia de Minkowski* que se define como:

$$M(x, y) = \left(\sum_i (|x_i - y_i|^p)^{\frac{1}{p}} \right)$$

Un último ejemplo de distancia entre dos variables x e y es la *distancia de Chebychev* que se define como:

$$C(x, y) = \max |x_i - y_i|$$

Entre las medidas de similitud (similaridad) tenemos los ya conocidos coeficientes de correlación de Pearson y Spearman y los múltiples coeficientes de asociación entre variables también conocidos (lambda, tau, etc.).

Para el caso de variables cualitativas, y en general para el caso de datos binarios (o dicotómicos), que son aquéllos que sólo pueden presentar dos opciones (blanco – negro, sí – no, hombre – mujer, verdadero – falso, etc.), existen diferentes medidas de proximidad o similitud, que se verán a continuación, partiendo de una tabla de frecuencias 2x2 en la que se representa el número de elementos de la población en los que se constata la presencia o ausencia del carácter (variable cualitativa) en estudio.

<i>Variable 1 →</i>		
<i>Variable 2 ↓</i>		
<i>Presencia</i>	<i>a</i>	<i>b</i>
<i>Ausencia</i>	<i>c</i>	<i>d</i>

Las principales medidas son las siguientes:

<i>Russel y Rao</i>	$RR_{xy} = \frac{a}{a+b+c}$	<i>Sokal y Sneath</i>	$SS_{xy} = \frac{2(a+d)}{2(a+d)+b+c}$
<i>Parejas simples</i>	$PS_{xy} = \frac{a+d}{a+b+c+d}$	<i>Rogers y Tanimoto</i>	$RT_{xy} = \frac{a+d}{a+d+2(b+c)}$
<i>Jaccard</i>	$J_{xy} = \frac{a}{a+b+c}$	<i>Sokal y Sneath(2)</i>	$SS2_{xy} = \frac{a}{a+2(b+c)}$
<i>Dice y Sorensen</i>	$D_{xy} = \frac{2a}{2a+b+c}$	<i>Kulczynski</i>	$K_{xy} = \frac{a}{b+c}$

Hay otro grupo de medidas denominadas medidas de similaridad para probabilidades condicionales, entre las que destacan las siguientes:

<i>Kulczynski (medida 2)</i>	$K2_{xy} = \frac{a/(a+b) + a/(a+c)}{2}$
<i>Sokal y Sneath (medida 4)</i>	$SS4_{xy} = \frac{a/(a+b) + a/(a+c) + d/(b+d) + d/(c+d)}{4}$
<i>Hamann</i>	$H_{xy} = \frac{(a+d)-(b+c)}{a+b+c+d}$

También suele considerarse un subgrupo de medidas denominadas de predicción entre las que se encuentran la D_{xy} de Anderberg, la Y_{xy} de Yule y la Q_{xy} de Yule, que se definen como sigue:

$$D_{xy} = \frac{\max(a, b) + \max(c, d) + \max(a, c) + \max(b, d) - \max(a + c, b + d) - \max(a + b, c + d)}{2(a + b + c + d)}$$

$$Y_{xy} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

$$Q_{xy} = \frac{ad - bc}{ad + bc}$$

Por último, se usan otras medidas binarias, entre las que destacan las siguientes:

<i>Ochiai</i>	$O_{xy} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$	<i>Sokal y Sneath (5)</i>	$SSS_{xy} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
<i>Sokal y Sneath (3)</i>	$SS3_{xy} = \frac{a+d}{b+c}$	<i>Correlación phi</i>	$\phi_{xy} = \frac{ad - bc}{(a+b)(a+c)(b+c)(c+d)}$
<i>Euclídea binaria</i>	$EB_{xy} = \sqrt{b+c}$	<i>Diferenciade forma</i>	$DF_{xy} = \frac{(a+b+c+d)(b+c) - (b-c)^2}{(a+b+c+d)^2}$
<i>Euclídea binaria</i> ²	$EB_{xy}^2 = b+c$	<i>Varianza disimilar</i>	$V_{xy} = \frac{b+c}{4(a+b+c+d)}$
<i>Dispersión</i>	$D_{xy} = \frac{ad - bc}{(a+b+c+d)^2}$	<i>Diferenciade tamaño</i>	$T_{xy} = \frac{(b-c)^2}{(a+b+c+d)^2}$
<i>Lancey Williams</i>	$LW_{xy} = \frac{b+c}{2a+b+c}$	<i>Diferenciade patrón</i>	$P_{xy} = \frac{bc}{(a+b+c+d)^2}$

CLUSTERS NO JERÁRQUICOS

La clasificación de todos los casos de una tabla de datos en grupos separados que configura el propio análisis proporciona clusters no jerárquicos. Esta denominación alude a la no existencia de una estructura vertical de dependencia entre los grupos formados y, por consiguiente, éstos no se presentan en distintos niveles de jerarquía. El análisis precisa que el investigador fije de antemano el número de clusters en que quiere agrupar sus datos.

Como puede no existir un número definido de grupos o, si existe, generalmente no se conoce, la prueba debe ser repetida con diferente número a fin de tantear la clasificación que mejor se ajuste al objetivo del problema, o la de más clara interpretación.

Los métodos no jerárquicos, también se conocen como *métodos partitivos* o de optimización, dado que, como hemos visto, tienen por objetivo realizar una sola partición de los individuos en K grupos. Esto implica que el investigador debe especificar a priori los grupos que deben ser formados. Ésta es, posiblemente, la principal diferencia respecto de los métodos jerárquicos. La asignación de individuos

a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia está en que estos métodos trabajan con la matriz de datos original y no requieren su conversión en una matriz de proximidades. Pedret agrupa los métodos no jerárquicos en las cuatro familias siguientes: *reasignación*, *búsqueda de la densidad*, *directos* y *reducción de dimensiones*.

Los *métodos de reasignación* permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior si esto optimiza el criterio de selección. El proceso termina cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Algunos de los algoritmos más conocidos dentro de estos métodos son el *método K-means* (*o K-medias*) de MeQueen (1967), el *Quick Cluster Analysis* y el *método de Forgy*, los cuales se suelen agrupar bajo el nombre de *métodos centroides o centros de gravedad*. Por otra parte está el *método de las nubes dinámicas*, debido a Diday.

Los *métodos de búsqueda de la densidad* presentan una aproximación tipológica y una aproximación probabilística. En la primera aproximación, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre los algoritmos más conocidos dentro de estos métodos están el *análisis modal de Wishart*, el *método de Taxmap de Carmichael y Sneath*, y el *método de Fortin*. En la segunda aproximación, se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Destaca en esta aproximación el *método de las combinaciones de Wolf*.

Los *métodos directos* permiten clasificar simultáneamente a los individuos y a las variables. Las entidades agrupadas, ya no son los individuos o las variables, sino que son las observaciones, es decir, los cruces que configuran la matriz de datos.

Los *métodos de reducción de dimensiones*, como el análisis factorial de tipo Q, guardan relación con el análisis cluster. Este método consiste en buscar factores en el espacio de los individuos, correspondiendo cada factor a un grupo. La interpretación de los grupos puede ser compleja dado que cada individuo puede corresponder a varios factores diferentes.

Resulta muy intuitivo suponer que una clasificación correcta debe ser aquélla en que la dispersión dentro de cada grupo formado sea la menor posible. Esta condición se denomina *criterio de varianza*, y lleva a seleccionar una configuración cuando la suma de las varianzas dentro de cada grupo (varianza residual) sea mínima.

Se han propuesto diversos algoritmos de clasificación no jerárquica, basados en minimizar progresivamente esta varianza, que difieren en la elección de los clusters provisionales que necesita el arranque del proceso y en el método de asignación de individuos a los grupos. Aquí se describen los dos más utilizados.

El **algoritmo de las H-medias** parte de una primera configuración arbitraria de grupos con su correspondiente media, eligiendo un primer individuo de arranque de cada grupo y asignando posteriormente cada caso al grupo cuya media es más cercana. Una vez que todos los casos han sido ubicados, calcula de nuevo las medias o centroides y las toma en lugar de los primeros individuos como una mejor aproximación de los mismos, repitiendo el proceso mientras la varianza residual vaya disminuyendo. La partición de arranque define el número de clusters que, lógicamente, puede disminuir si ningún caso es asignado a alguno de ellos.

El **algoritmo de las K-medias**, el más importante desde los puntos de vista conceptual y práctico, parte también de unas medias arbitrarias y, mediante pruebas sucesivas, contrasta el efecto que sobre la varianza residual tiene la asignación de cada uno de los casos a cada uno de los grupos. El valor mínimo de varianza determina una configuración de nuevos grupos con sus respectivas medias. Se asignan otra vez todos los casos a estos nuevos centroides en un proceso que se repite hasta que ninguna transferencia puede ya disminuir la varianza residual; o se alcance otro criterio de parada: un número limitado de pasos de iteración o, simplemente, que la diferencia obtenida entre los centroides de dos pasos consecutivos sea menor que un valor prefijado. El procedimiento configura los grupos maximizando, a su vez, la distancia entre sus centros de gravedad. Como la varianza total es fija, minimizar la residual hace máxima la factorial o intergrupos. Y puesto que minimizar la varianza residual es equivalente a conseguir que sea mínima la suma de distancias al cuadrado desde los casos a la media del cluster al que van a ser asignados, es esta distancia euclídea al cuadrado la utilizada por el método.

Como se comprueban los casos secuencialmente para ver su influencia individual, el cálculo puede verse afectado por el orden de los mismos en la tabla; pese a lo cual es el algoritmo que mejores resultados produce. Otras variantes propuestas a este método llevan a clasificaciones muy similares.

Como cualquier otro método de clasificación no jerárquica, proporciona una solución final única para el número de clusters elegido, a la que se llegará con menor número de iteraciones cuanto más cerca estén las “medias” de arranque de las que van a ser finalmente obtenidas. Los programas automáticos seleccionan generalmente estos primeros valores, tantos como grupos se pretenda formar, entre los puntos más separados de la nube.

Los clusters no jerárquicos están indicados para grandes tablas de datos, y son también útiles para la detección de casos atípicos: Si se elige previamente un número elevado de grupos, superior al deseado, aquéllos que contengan muy escaso número de individuos servirían para detectar casos extremos que podrían distorsionar la configuración. Es aconsejable realizar el análisis definitivo sin ellos, ya con el número deseado de grupos para después, opcionalmente, asignar los atípicos al

cluster adecuado que habrá sido formado sin su influencia distorsionante. Un problema importante que tiene el investigador para clasificar sus datos en grupos es, como se ha dicho, la elección de un número adecuado de clusters. Puesto que siempre será conveniente efectuar varios tanteos, la selección del más apropiado al fenómeno que se estudia ha de basarse en criterios tanto matemáticos como de interpretabilidad. Entre los primeros, se han definido numerosos indicadores de adecuación como el Criterio cúbico de clusters y la Pseudo F que se describen en el ejemplo de aplicación práctica. El uso inteligente de estos criterios, combinado con la interpretabilidad práctica de los grupos, constituye el arte de la decisión en la clasificación multivariante de datos.

Matemáticamente, un método de clasificación no jerarquizado consiste en formar un número prefijado K de clases homogéneas excluyentes, pero con máxima divergencia entre las clases. Las K clases o clusters forman una única partición (*clustering*) y no están organizadas jerárquicamente ni relacionadas entre sí. La clasificación no jerárquica o de reagrupamiento tiene una estructura matemática menos precisa que la clasificación jerárquica. El número de métodos existentes ha crecido excesivamente en los últimos años y algunos problemas derivados de su utilización todavía no han sido resueltos.

Supongamos que N es el número de sujetos a clasificar formando K grupos, respecto a n variables X_1, \dots, X_n . Sean W , B y T las matrices de dispersión dentro grupos, entre grupos y total respectivamente. Como $T = B + W$ y T no depende de la forma en que han sido agrupados los sujetos, un criterio razonable de clasificación consiste en construir K grupos de forma que B sea máxima o W sea mínima, siguiendo algún criterio apropiado. Algunos de estos criterios son:

- a) Minimizar $\text{Traza}(W)$
- b) Minimizar $\text{Determinante}(W)$
- c) Minimizar $\text{Det}(W)/\text{Det}(T)$
- d) Maximizar $\text{Traza}(W^T B)$
- e) Minimizar $\sum_{i=1}^K \sum_{h=1}^{N_i} (X_{ih} - \bar{X}_i)' S_i^{-1} (X_{ih} - \bar{X}_i)$

Los criterios a) y b) se justifican porque tratan de minimizar la magnitud de la matriz W . El criterio e) es llamado *criterio de Wilks* y es equivalente a b) porque $\det(T)$ es constante. El caso d) es el llamado *criterio de Hottelling* y el criterio e) representa la suma de las distancias de Mahalanobis de cada sujeto al centroide del grupo al que es asignado.

Como el número de formas de agrupar N sujetos en K grupos es del orden de $k^N * k!$, una vez elegido el criterio de optimización, es necesario seguir algún algoritmo adecuado de clasificación para evitar un número tan elevado de agrupamientos.

El método ISODATA, introducido por Ball y Hall (1967), es uno de los más conocidos. Esencialmente consiste en partir de K clases (construidas por ejemplo aleatoriamente) y reasignar un sujeto de una clase i a una clase j si se mejora el criterio elegido de optimización. Para un seguimiento matemático de estos métodos véase Gnanadesikan (1977) y Escudero (1977).

CLUSTERS JERÁRQUICOS: DENDOGRAMA

Es frecuente en la investigación biológica la necesidad de clasificar los datos en grupos con estructura arborescente de dependencia, de acuerdo con diferentes niveles de jerarquía. La clasificación de especies animales o vegetales constituye un buen ejemplo de este interés científico. Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clusters los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al cluster final que contiene todos los casos analizados. La representación gráfica de estas etapas (Figura 13-1) de formación de grupos, a modo de árbol invertido, se denomina *dendograma*.

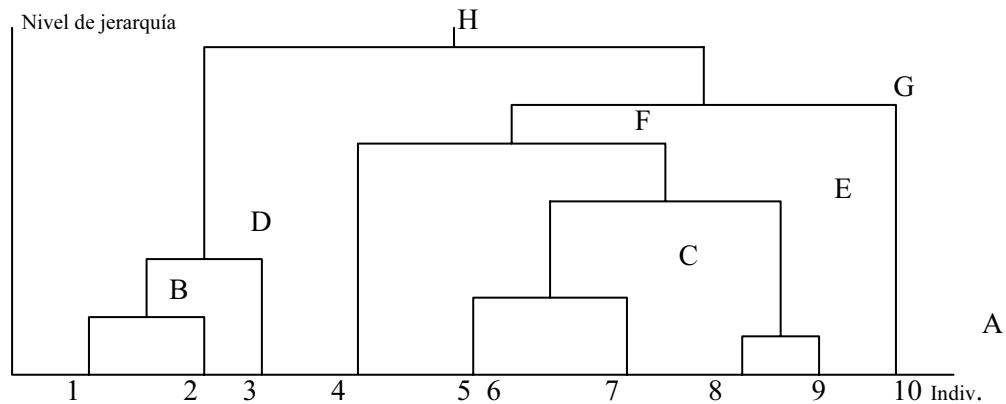


Figura 11-1

La figura, que corresponde a un estudio de los individuos, muestra cómo el 8 y el 9 se agrupan en un primer cluster (A). En un nivel inmediatamente superior, se unen los individuos 1 y 2 (cluster B); y enseguida los 5, 6, y 7 (C). Un paso siguiente engloba el cluster B con el individuo 3 (D); y así sucesivamente hasta que todos ellos quedan estructurados al conseguir, en el nivel más alto, el cluster total (H) que reúne los 10 casos.

Evidentemente, la decisión de todas estas agrupaciones ha de tomarse en función de la *similaridad multivariante* (o de su contrario *distancia*) proporcionada por el conjunto de variables estudiadas, ya que en cada nivel de jerarquía se unen los dos clusters más cercanos. Es, pues, importante como paso previo a un análisis de clusters jerárquicos, la elección de una adecuada métrica de similaridad o disimilaridad. Se sabe que a partir de la tabla inicial de datos ($n \times p$) es preciso calcular una matriz de *distancias* entre individuos ($n \times n$). Este concepto, ya introducido anteriormente, merece aquí un tratamiento más detallado. Se han descrito numerosas formas de medir distancias multivariantes. La conocida distancia euclídea es la más sencilla y utilizada: Se usa también en el análisis de componentes principales cuyos factores son, como se sabe, muchas veces datos previos para entrar en un análisis de clusters.

Para variables cualitativas puede emplearse la distancia χ^2 , y, si son dicotómicas, la distancia de Jaccard. La distancia euclídea al cuadrado, la euclídea generalizada, la de bloques o Manhattan, la de Tchebycheff, la de Mahalanobis, y otras medidas de similaridad como los coeficientes de correlación de Pearson y de correlación por rangos de Kendall entre individuos, el índice de Gower, etc. dan idea de la enorme variedad de formas de enfocar el diseño de un análisis de clasificación de datos, cada una de ellas con sus ventajas e inconvenientes que, en definitiva, serán mejores o peores según las características del fenómeno estudiado y, sobre todo, de la relevancia o interpretabilidad de los grupos obtenidos. Sin embargo, las distancias más usadas son pocas y ya han sido definidas.

La segunda decisión que el investigador debe tomar es, precisamente, qué algoritmo emplear para la formación de grupos, definiendo a qué va a llamar “distancia entre clusters” para luego poder unir, a otro nivel jerárquico, aquellos clusters más próximos. Este concepto no existía en el análisis no jerárquico, puesto que allí no se unían los grupos. Es también muy variada, y continuamente ampliada, la oferta de procedimientos de agrupación.

La aglomeración comienza con tantos grupos como individuos; cada uno de éstos constituye un cluster inicial. A medida que transcurren las etapas del proceso se van formando nuevos clusters por unión de dos individuos, de un individuo con un grupo previo, o de dos grupos anteriores entre los que exista la menor distancia.

El proceso finaliza con un único grupo (todos los individuos), pero constituido por aglomeraciones sucesivas en distintos niveles. Este es el fundamento de la agregación (ascendente); en contraposición con el proceso de disagregación (descendente), que opera de forma inversa: Parte del grupo total de individuos para llegar, tras varias etapas de partición, hasta tantos clusters como individuos. Característica importante de los métodos jerárquicos es el no permitir reasignaciones de grupos, es decir, que dos clusters (o dos individuos) que han sido unidos en un paso del proceso no pueden ya separarse en etapas sucesivas; lo que sí es posible en los métodos no jerárquicos (si bien en éstos es necesario, como se ha visto, fijar previamente el número de clusters deseado).

Se presenta a continuación una relación de los principales métodos de unión de grupos o algoritmos de clasificación jerárquica en la que, junto a su descripción, se comentan sus ventajas e inconvenientes como ayuda en la decisión del método apropiado. Suele distinguirse entre métodos aglomerativos y métodos disociativos. Entre los **métodos aglomerativos** tenemos los siguientes:

Método de las distancias mínimas o Enlace simple (single linkage). Considera como distancia entre dos grupos la que responde al concepto de “vecinos más cercanos” (*nearest neighbor*), es decir, la separación que existe entre los individuos más próximos de uno y otro grupo. Aunque presenta buenas propiedades teóricas, su eficacia no ha sido la esperada en estudios de validación comparativa de métodos mediante ficheros simulados (Monte Carlo), por su tendencia al “encadenamiento” (une clusters realmente diferentes, si están próximos), y porque sólo considera la información de individuos extremos (los valores atípicos, *outliers*, pueden distorsionar la agrupación). Si bien no es adecuado para la obtención de grupos compactos, resulta de utilidad para clusters irregulares o elongados. El método consiste en ir agrupando los individuos que tienen menor distancia o mayor similitud, considerando como distancia entre dos clusters la distancia entre sus dos puntos más próximos.

Método de las distancias máximas o Enlace completo (complete linkage). Considera como distancia entre dos grupos la existente entre “vecinos más lejanos” (*furthest neighbor*), es decir, entre los individuos más separados de ambos grupos (máxima distancia que es posible encontrar entre un caso de un cluster y un caso de otro). Presenta una excesiva tendencia a producir grupos de igual diámetro, y se ve muy distorsionado ante valores atípicos moderados.

Método del promedio entre grupos o Enlace promedio (average linkage). Considera como distancia entre dos clusters, no la de los individuos más próximos ni más lejanos de ambos grupos, sino la distancia media entre todos los pares posibles de casos (uno de cada cluster). Tiende a producir clusters compactos, por lo que es muy utilizado y suele ser el método por defecto en los paquetes de software. Una variante de este método es el **método de la media ponderada (average linkage within groups)**, en el cual se combinan los grupos de tal forma que la distancia promedio entre todos los casos en el cluster resultante sea lo más pequeña posible.

Método del centroide o Enlace centroide (centrod method). Considera como distancia entre dos grupos la existente entre sus centros de gravedad, definidos por las medias aritméticas de las variables de los individuos que componen los clusters. Es el más robusto de los métodos jerárquicos ante la presencia de casos atípicos.

Método de la mediana (median method). Considera como distancia entre dos grupos la existente entre las medianas de las variables de los individuos que componen los clusters. De este modo, los dos clusters que se combinan se ponderan de forma equivalente al método centroide, pero independientemente del número de individuos que haya en cada grupo.

Método de Ward o Enlace por mínima varianza (momento central de orden dos o pérdida de inercia mínima). Considera como distancia entre dos grupos el menor incremento de varianza residual global, o sea, si en un nivel dado existe un número de clusters de los que se deben elegir dos para una nueva fusión, se prueban todas las parejas posibles y se calcula la varianza residual global o intragrupo con cada pareja unida y todos los demás clusters. La pareja de grupos que produzca el mínimo incremento en esta varianza residual será la elegida para su unión en un nuevo nivel. En el último nivel, con todos los individuos agrupados en un sólo cluster, la varianza residual es máxima y coincide con la varianza total al ser, lógicamente, nula la varianza factorial o intergrupos (ya no hay grupos). Tiende a formar clusters esféricos o compactos, y del mismo tamaño. Requiere una distribución normal multivariante en las variables del estudio. Utiliza más información sobre el contenido de los grupos que otros métodos. Es, junto con el enlace promedio, el que ha demostrado mayor eficacia en estudios de simulación.

Siendo más precisos, en el método de Ward se calcula la media de todas las variables de cada cluster, luego se calcula la distancia euclídea al cuadrado entre cada individuo y la media de su grupo y después se suman las distancias de todos los casos. En cada paso, los clusters que se forman son aquéllos que resultan con el menor incremento en la suma total de las distancias al cuadrado intracumple. La métrica normalmente considerada en los métodos hasta aquí descritos es la euclídea o la euclídea al cuadrado. Esta última se suele usar por omisión en programas estadísticos.

Método del Enlace por densidad. Utiliza un nuevo concepto de distancia entre dos individuos (A y B): Puesto que la nube de puntos no tendrá una densidad homogénea, existirán zonas de acúmulo de casos separadas por otras menos densas, lo que puede sugerir agrupamientos naturales entre los individuos. De esta forma, para unir dos de ellos, se les puede considerar como centros de dos esferas (multidimensionales) capaces de englobar un número prefijado de casos de la nube (por ejemplo, $k = 4$). Se define la densidad de cada esfera como su “masa”, o número de individuos que contiene (en frecuencia relativa), dividida por su volumen. Como el numerador es constante (k/n) y en el denominador figura el radio, para abarcar 4 casos muy próximos, el radio necesario será pequeño y la densidad grande. Es lógico, por tanto, definir conceptualmente una distancia como el inverso de una densidad. Al ser dos las esferas consideradas, se toma como distancia entre A y B la media de los inversos de ambas densidades. Así, usando esta distancia, se ponderan más próximos dos individuos, cada uno de ellos con k casos muy cercanos, que otros dos con k casos lejanos. En la Figura 13-2, el individuo A sería unido en el siguiente nivel al B, y no al C, a pesar de que existe la misma separación geométrica entre AB y AC. Esferas que no se cortan hacen no considerar a sus individuos centrales como candidatos a ser unidos. Ello indicará que no están suficientemente «próximos», o que el número k de casos a abarcar debe ser mayor. Para impedir esta unión, por convenio se considera una distancia infinita cuando las esferas no se cortan; por ello, no siempre con este método se llega al final de la aglomeración a un cluster único. La sucesiva fusión de éstos se realiza mediante el método del enlace simple, pero con esta distancia definida en

lugar de la euclídea. El enlace por densidad es capaz de encontrar clusters irregulares y elongados, aunque ha demostrado menos eficacia en la configuración de clusters compactos. Se han descrito diversas técnicas basadas en este algoritmo de clasificación, como por ejemplo el enlace por densidad.

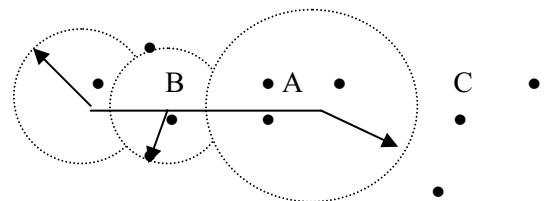


Figura 11-2

Al aplicar cualquiera de estos métodos existe la posibilidad de que, en un determinado paso de unión, se produzcan «empates» en las distancias de dos o más individuos o grupos. En ese nivel jerárquico podría, pues, agregarse cualquiera de ellos, y se hace necesario elegir uno. Esta decisión es tomada automáticamente por los programas de acuerdo con un criterio arbitrario pero definido (por ejemplo, basado en el orden que ocupan los casos en la tabla de datos original). Si estos empates se presentan en los niveles más bajos del dendograma afectarán poco a la estructura, pero si aparecen en niveles altos podría variar sustancialmente la configuración final según la decisión tomada por el programa. Una comprobación del grado de afectación podría hacerse, por ejemplo, permutando el orden de los individuos en la tabla. Altas precisiones en las medidas de las variables alejan este problema de empates.

Fórmula de Lance y Williams para la distancia entre grupos

Matemáticamente, Lance y Williams desarrollaron una fórmula general que puede ser utilizada para describir los distintos tipos de enlaces de los métodos jerárquicos aglomerativos. La **fórmula de Lance y Williams para la distancia entre grupos** es la siguiente:-

$$D_{k(i,j)} = \alpha_i D_{ki} + \alpha_j D_{kj} + \beta D_{ij} + \gamma |D_{ki} - D_{kj}|$$

donde D_{ij} es la distancia entre los grupos i y j , y α , β y γ son los tres parámetros del modelo. Se observa lo siguiente:

$$\alpha_i = \alpha_j = 1/2, \beta = 0 \text{ y } \gamma = -1/2 \Rightarrow \text{enlace simple}$$

$$\alpha_i = \alpha_j = 1/2, \beta = 0 \text{ y } \gamma = 1/2 \Rightarrow \text{enlace completo}$$

$$\alpha_i = \alpha_j = 1/2, \beta = -1/4 \text{ y } \gamma = 0 \Rightarrow \text{método de la mediana}$$

$$\alpha_i = \frac{n_i}{n_i + n_j}, \alpha_j = \frac{n_j}{n_i + n_j}, \beta = -\alpha_i \alpha_j \text{ y } \gamma = 0 \Rightarrow \text{enlace centroide}$$

$$\alpha_i = \frac{n_i}{n_i + n_j}, \alpha_j = \frac{n_j}{n_i + n_j}, \beta = \gamma = 0 \Rightarrow \text{enlace promedio}$$

$$\alpha_i = \frac{n_k + n_i}{n_k + n_i + n_j}, \alpha_j = \frac{n_k + n_j}{n_k + n_i + n_j}, \beta = \frac{-n_k}{n_k + n_i + n_j} \text{ y } \gamma = 0 \Rightarrow \text{Ward}$$

$\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, $\beta < 1$ y $\gamma = 0 \Rightarrow$ método flexible (cuádruple restricción)

El último método (**cuádruple restricción**) consiste en utilizar la forma de Lance y Williams variando los coeficientes según las necesidades del clasificador, pero respetando las cuatro restricciones impuestas.

Los métodos de clusters jerárquicos, por la laboriosidad de los cálculos, no resultan prácticos para procesar grandes ficheros de datos. En estos casos, puede ser aconsejable realizar un análisis previo no jerárquico, que proporcione un número preliminar razonable de clusters (en lugar de individuos) que servirán luego de partida para su posterior clasificación jerárquica.

Como resumen, los métodos jerárquicos producen resultados más ricos que los no jerárquicos. Con un solo análisis se obtiene una configuración de grupos en cada nivel de clasificación. Los mismos indicadores que en clasificación no jerárquica valoraban la adecuación del número de clusters (Criterio cúbico de clusters, Pseudo F, etc.) permiten detectar aquí el nivel jerárquico en que la separación de los grupos formados es más ostensible.

Los siete criterios que se acaban de exponer para la clasificación jerárquica aglomerativa también son utilizables en el caso de los **métodos de clasificación disociativos**, si bien en la práctica los que más suelen usarse son el del promedio entregrupos y el de Ward. Además de los anteriores, dentro del proceso disociativo, destacan los que se presentan muy brevemente a continuación.

En los métodos disociativos, cuando el criterio de división toma en consideración cada variable observada una a una, el método recibirá el nombre de **monotético**. Por el contrario, cuando se toman en cuenta todas las variables simultáneamente el método se llamarán **polítetico**. El análisis de asociación está especialmente diseñado para el caso de variables dicotómicas. Según el **método asociativo de Williams y Lambert** se construyen tablas de contingencia 2 x 2, para cada par de variables y se calcula la ji-cuadrado para cada tabla. El criterio de participación en los grupos se basa en la variable que maximiza la ji-cuadrado.

Otro método disociativo importante es el **detector automático de interacción**, conocido como **AID (Automatic Interaction Detector Method)**, que no nació como una técnica de clasificación y sin embargo, ha sido aceptada como tal. Los elementos básicos del AID son análogos a los de la regresión, con una variable dependiente y varias independientes. El objetivo del AID consiste en determinar qué variables independientes proporcionan la mayor diferencia a las distintas medias de la variable dependiente para los diferentes grupos.

Los dos métodos anteriores son monotéticos. Como ejemplo de técnica politéctica, Sierra expone la siguiente: El grupo inicial de individuos se divide en dos, separando 'uno a uno' los elementos mediante el siguiente criterio. 1) se separa el individuo cuya distancia media al resto de los individuos sea mayor para formar el grupo A; 2) se estudia qué elemento del grupo original es el que se separa más para ir a formar parte del grupo A; 3) este proceso se repite hasta que los individuos que queden estén más próximos del grupo original que del A; 4) se repite el proceso para cada uno de los subgrupos que secuencialmente se vayan obteniendo.

ANÁLISIS DE CONGLOMERADOS EN DOS FASES

En algunas aplicaciones, se puede seleccionar como método el *Análisis de conglomerados en dos fases*. Ofrece una serie de funciones únicas que se detallan a continuación:

- Selección automática del número más apropiado de conglomerados y medidas para la selección de los distintos modelos de conglomerado.
- Posibilidad de crear modelos de conglomerado basados al mismo tiempo en variables categóricas y continuas.
- Posibilidad de guardar el modelo de conglomerados en un archivo XML externo y, a continuación, leer el archivo y actualizar el modelo de conglomerados con datos más recientes.
- Asimismo, el *Análisis de conglomerados en dos fases* puede analizar archivos de datos grandes.

CLASIFICACIÓN Y SEGMENTACIÓN MEDIANTE ANÁLISIS CLUSTER CON SPSS

PRINCIPIOS DEL ANÁLISIS CLUSTER

El análisis cluster, cuyo esquema se muestra en la Figura 14-1, es una técnica exploratoria de análisis estadístico de datos diseñada para revelar concentraciones en los datos o en las variables y que sugiere modos potencialmente útiles de agrupar las observaciones. Es muy importante tener presente que pueden agruparse tanto casos como variables. El análisis cluster o de conglomerados divide las observaciones en grupos basándose en la proximidad o lejanía de unas con otras, por lo tanto es esencial el uso adecuado del concepto de distancia. Las observaciones muy cercanas deben de caer dentro del mismo cluster y las muy lejanas deben de caer en clusters diferentes, de modo que las observaciones dentro de un cluster sean homogéneas y lo más diferentes posibles de las contenidas en otros clusters.

También hay que tener muy presente el tipo de datos que se maneja. Si las variables de aglomeración están en escalas completamente diferentes será necesario estandarizar previamente las variables, o por lo menos trabajar con desviaciones respecto de la media. Es necesario observar también los valores atípicos y desaparecidos porque los métodos jerárquicos no tienen solución con valores perdidos y los valores atípicos deforman las distancias y producen clusters unitarios. También es nocivo para el análisis cluster la presencia de variables correlacionadas, de ahí la importancia del análisis previo de multicolinealidad. Si es necesario se realiza un análisis factorial previo y posteriormente se aglomeran las puntuaciones.

La solución del análisis cluster no tiene porqué ser única, pero no deben encontrarse soluciones contradictorias por distintos métodos. El número de observaciones en cada cluster debe ser relevante, ya que en caso contrario puede haber valores atípicos. Además, los conglomerados deben de tener sentido conceptual y no variar mucho al variar la muestra o el método de aglomeración.

ESQUEMA GENERAL DEL ANÁLISIS CLUSTER

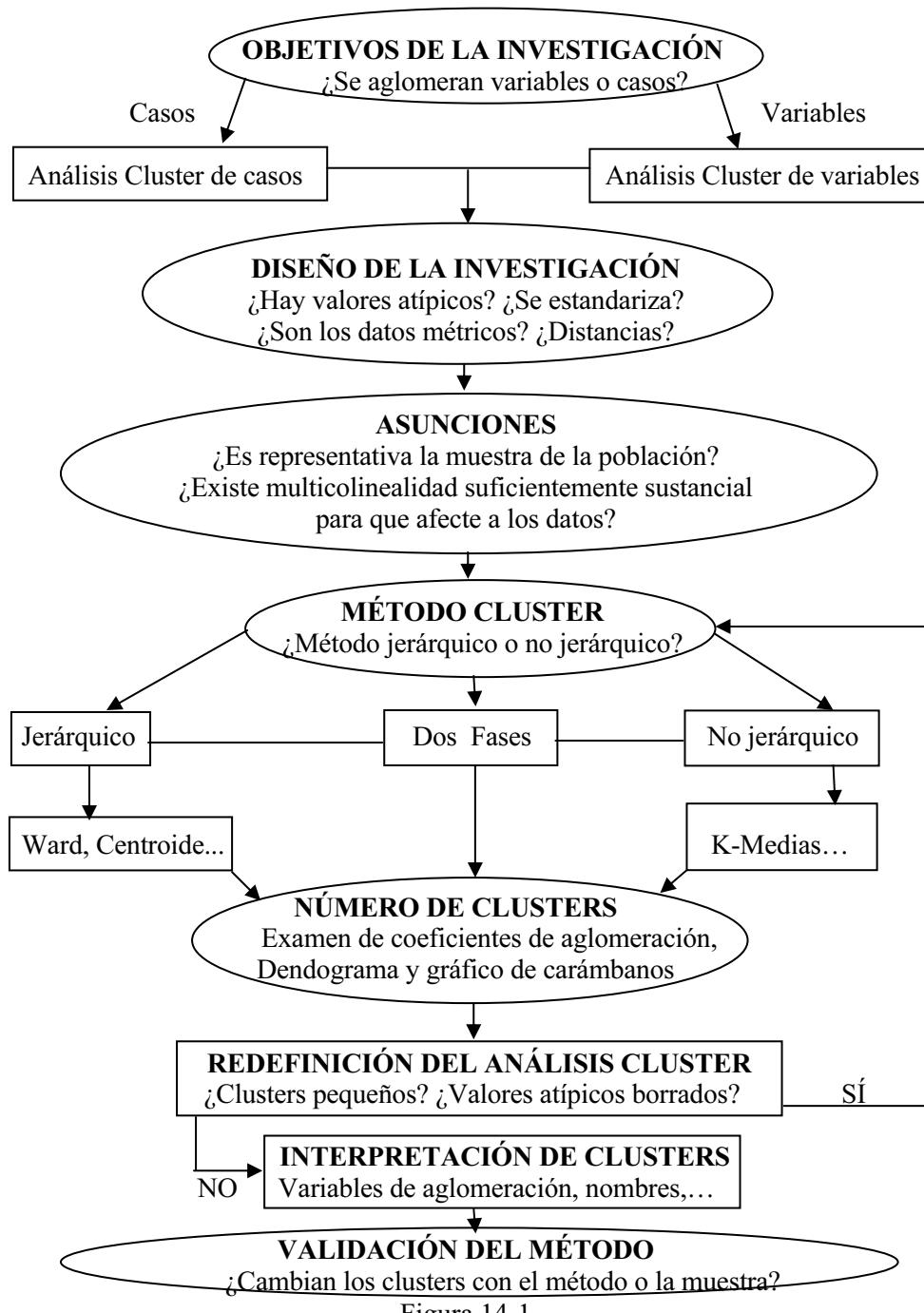


Figura 14-1

SPSS Y EL ANÁLISIS CLUSTER NO JERÁRQUICO

SPSS incorpora un procedimiento que implementa el análisis cluster no jerárquico mediante el método k-medias. Este procedimiento intenta identificar grupos de casos relativamente homogéneos basándose en las características seleccionadas y utilizando un algoritmo que puede gestionar un gran número de casos. Sin embargo, el algoritmo requiere que el usuario especifique el número de conglomerados. Puede especificar los centros iniciales de los conglomerados si conoce de antemano dicha información. Puede elegir uno de los dos métodos disponibles para clasificar los casos: la actualización de los centros de los conglomerados de forma iterativa o sólo la clasificación. Asimismo, puede guardar la pertenencia a los conglomerados, información de la distancia y los centros de los conglomerados finales. Si lo desea, puede especificar una variable cuyos valores sean utilizados para etiquetar los resultados por casos. También puede solicitar los estadísticos F de los análisis de varianza. Aunque estos estadísticos son oportunistas (ya que el procedimiento trata de formar grupos que de hecho difieran), el tamaño relativo de los estadísticos proporciona información acerca de la contribución de cada variable a la separación de los grupos.

Para la solución completa se obtendrán los centros iniciales de los conglomerados y la tabla de ANOVA. Para cada caso se obtendrá información del conglomerado y la distancia desde el centro del conglomerado.

Como ejemplo podemos preguntar: ¿Cuáles son los grupos identificables de países con población, densidad de población y población urbana similares? Con el análisis de conglomerados de k-medias, podrían agruparse los países en k grupos homogéneos basados en las características consideradas.

Para realizar un análisis cluster no jerárquico de k-medias, elija en los menús *Anализar → Clasificar → Conglomerado de k medias* (Figura 14-2) y seleccione las variables y las especificaciones para el análisis (Figura 14-3). Previamente es necesario cargar en memoria el fichero de nombre MUNDO mediante *Archivo → Abrir → Datos*. Este fichero contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo. Las variables clasificadoras a considerar son: población (*poblac*), la población urbana (*urbana*) y densidad de (*densidad*). Como variable de agrupación usamos el país (*país*).

En cuanto a los datos, las variables deben ser cuantitativas en el nivel de intervalo o de razón. Si las variables son binarias o recuentos, utilice el procedimiento *Análisis de conglomerados jerárquicos*.



Figura 14-2

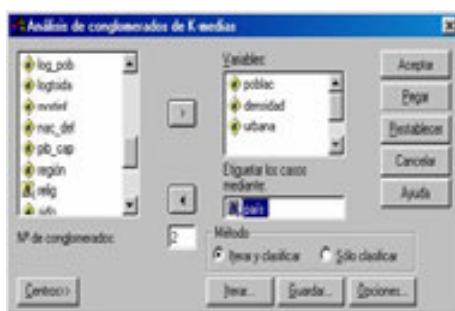


Figura 14-3

Las distancias se calculan utilizando la distancia euclídea simple. Si desea utilizar otra medida de distancia o de similaridad, utilice el procedimiento *Análisis de conglomerados jerárquicos*. El escalamiento de las variables es una consideración importante. Si sus variables utilizan diferentes escalas (por ejemplo, una variable se expresa en dólares y otra en años), los resultados podrían ser equívocos. En estos casos, debería considerar la estandarización de las variables antes de realizar el análisis de conglomerados de k-medias (esto se puede hacer en el procedimiento *Descriptivos*). Este procedimiento supone que ha seleccionado el número apropiado de conglomerados y que ha incluido todas las variables relevantes. Si ha seleccionado un número inapropiado de conglomerados o ha omitido variables relevantes, los resultados podrían ser equívocos.

El botón *Opciones* de la Figura 14-3 nos lleva a la pantalla de la Figura 14-4, en cuyo cuadro *Estadísticos* se establecen los estadísticos más relevantes relativos a las variables que ofrecerá el análisis, que son: centros de conglomerados iniciales, tabla de ANOVA e información del conglomerado para cada caso. En el cuadro *Valores perdidos* se elige la forma de su exclusión. Las opciones disponibles son: excluir casos según lista o excluir casos según pareja.

El botón *Iterar* (sólo disponible si se ha seleccionado el método *Iterar y clasificar* en el cuadro de diálogo principal de la Figura 14-3) nos lleva a la pantalla de la Figura 14-5 cuya opción *Nº máximo de iteraciones* limita el número de iteraciones en el algoritmo k-medias, de modo que el proceso iterativo se detiene después de este número de iteraciones, incluso si no se ha satisfecho el criterio de convergencia. Este número debe estar entre el 1 y el 999. Para reproducir el algoritmo utilizado por el comando Quick Cluster en las versiones previas a la 5.0, establezca a 1 el número máximo de iteraciones. La opción *Criterio de convergencia* determina cuándo cesa la iteración y representa una proporción de la distancia mínima entre los centros iniciales de los conglomerados, por lo que debe ser mayor que 0 pero no mayor que 1. Por ejemplo, si el criterio es igual a 0,02, la iteración cesará si una iteración completa no mueve ninguno de los centros de los conglomerados en una distancia superior al dos por ciento de la distancia menor entre cualquiera de los centros iniciales.

La opción *Usar medias actualizadas* permite solicitar la actualización de los centros de los conglomerados tras la asignación de cada caso. Si no selecciona esta opción, los nuevos centros de los conglomerados se calcularán después de la asignación de todos los casos.

El botón *Guardar* permite guardar información sobre la solución como nuevas variables para que puedan ser utilizadas en análisis subsiguientes. Estas variables son: *Conglomerado de pertenencia*, que crea una nueva variable que indica el conglomerado final al que pertenece cada caso (los valores de la nueva variable van desde el 1 hasta el número de conglomerados) y *Distancia desde centro del conglomerado*, que indica la distancia euclídea entre cada caso y su centro de clasificación.

El botón *Centros* permite al usuario especificar sus propios centros iniciales para los conglomerados (*Leer iniciales de*) o guardar los centros finales para análisis subsiguientes (*Guardar finales en*).

El botón *Pegar* genera la sintaxis del comando a partir de las selecciones del cuadro de diálogo y pega dicha sintaxis en la ventana de sintaxis designada. Para poder pulsar en *Pegar*, debe seleccionar al menos una variable.

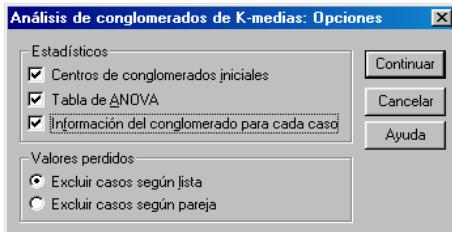


Figura 14-4

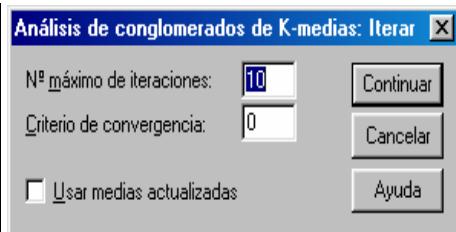


Figura 14-5

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 14-3 para obtener los resultados del análisis cluster de k-medias según se muestra en la Figura 14-6. En la parte izquierda de la Figura podremos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. A continuación, se presentan los centros iniciales de los conglomerados y el historial de iteraciones. En la Figura 14-7 se presentan los centros de los conglomerados finales y el número de casos en cada conglomerado.

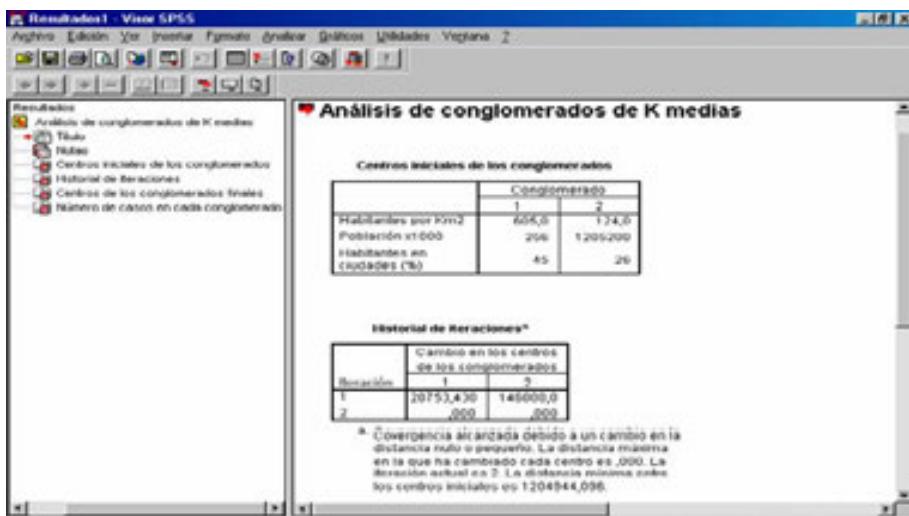


Figura 14-6

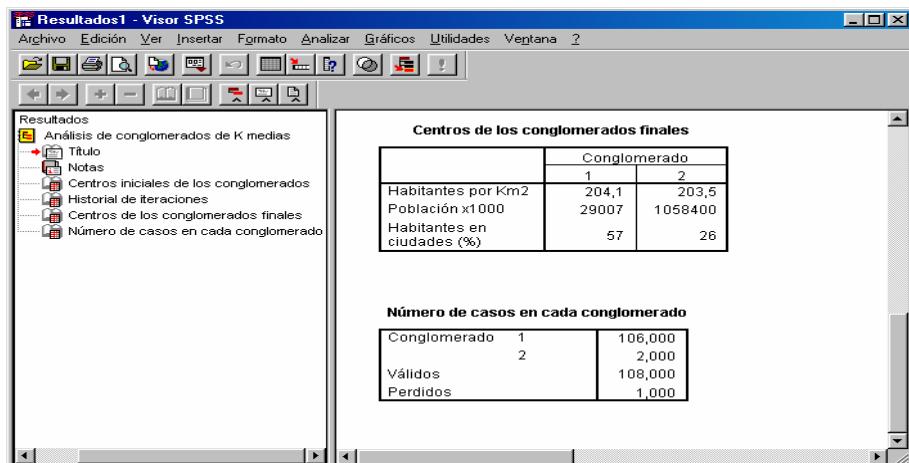


Figura 14-7

SPSS Y EL ANÁLISIS CLUSTER JERÁRQUICO

SPSS incorpora un procedimiento que implementa el análisis cluster no jerárquico. Este procedimiento intenta identificar grupos relativamente homogéneos de casos (o de variables) basándose en las características seleccionadas, mediante un algoritmo que comienza con cada caso (o cada variable) en un conglomerado diferente y combina los conglomerados hasta que sólo queda uno. Es posible analizar las variables brutas o elegir de entre una variedad de transformaciones de estandarización. Las medidas de distancia o similaridad se generan mediante el procedimiento Proximidades. Los estadísticos se muestran en cada etapa para ayudar a seleccionar la mejor solución.

Como ejemplo podemos preguntar: ¿Cuáles son los grupos identificables de países con población, densidad de población, población urbana y esperanza de vida femenina similares? Durante el proceso se obtiene el historial de conglomeración, la matriz de distancias (o similaridades) y la pertenencia a los conglomerados para una solución única o una serie de soluciones. También se obtienen dendrogramas y diagramas de témpanos.

Para realizar un análisis cluster jerárquico, elija en los menús *Analizar* → *Clasificar* → *Conglomerados jerárquicos* (Figura 14-8) y seleccione las variables y las especificaciones para el análisis (Figura 14-9). Previamente es necesario cargar en memoria el fichero de nombre MUNDO mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo. Las variables clasificadoras a considerar son: población (*poblac*), la población urbana (*urbana*), la densidad de (*densidad*) y la esperanza de vida femenina (*espvidaf*). Como variable de agrupación usamos el país (*país*). Si las variables son binarias o recuentos, se utiliza sólo el procedimiento *Análisis de conglomerados jerárquicos*.

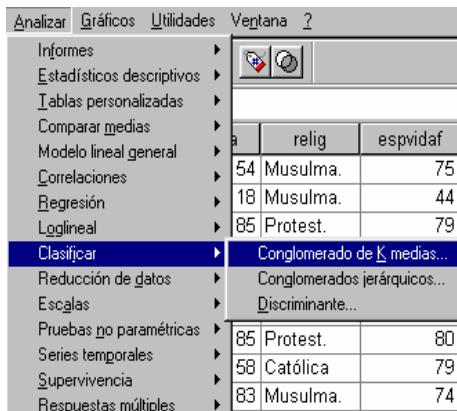


Figura 14-8



Figura 14-9

El botón *Estadísticos* de la Figura 14-9 nos lleva a la pantalla de la Figura 14-10, cuya opción *Historial de conglomeración* muestra los casos o conglomerados combinados en cada etapa, las distancias entre los casos o los conglomerados que se combinan, así como el último nivel del proceso de aglomeración en el que cada caso (o variable) se unió a su conglomerado correspondiente. La opción *Matriz de distancias* proporciona las distancias o similaridades entre los elementos. El campo *Conglomerado de pertenencia* muestra el conglomerado al cual se asigna cada caso en una o varias etapas de la combinación de los conglomerados. Las opciones disponibles son: *Solución única* y *Rango de soluciones*.

El botón *Método* de la Figura 14-9 nos lleva a la Figura 14-11, cuya opción *Método de conglomeración* permite elegir dicho método. Las opciones disponibles son: *Vinculación inter-grupos*, *Vinculación intra-grupos*, *Vecino más próximo*, *Vecino más lejano*, *Agrupación de centroides*, *Agrupación de medianas* y *Método de Ward*. El cuadro *Medida* de la Figura 14-11 permite especificar la medida de distancia o similaridad que será empleada en la aglomeración. Seleccione el tipo de datos y la medida de distancia o similaridad adecuada. En la opción *Intervalo* (Figura 14-12) las opciones disponibles son: *Distancia euclídea*, *Distancia euclídea al cuadrado*, *Coseno*, *Correlación de Pearson*, *Chebychev*, *Bloque*, *Minkowski* y *Personalizada*. En la opción *Datos de frecuencias* las opciones disponibles son: *Medida de Chi-cuadrado* y *Medida de Phi-cuadrado*. En la opción *Datos binarios* (Figura 14-13) las opciones disponibles son: *Distancia euclídea*, *Distancia euclídea al cuadrado*, *Diferencia de tamaño*, *Diferencia de configuración*, *Varianza*, *Dispersión*, *Forma*, *Concordancia simple*, *Correlación Phi de 4 puntos*, *Lambda*, *D de Anderberg*, *Dice*, *Hamann*, *Jaccard*, *Kulczynski 1*, *Kulczynski 2*, *Lance y Williams*, *Ochiai*, *Rogers y Tanimoto*, *Russel y Rao*, *Sokal y Sneath 1*, *Sokal y Sneath 2*, *Sokal y Sneath 3*, *Sokal y Sneath 4*, *Sokal y Sneath 5*, *Y de Yule* y *Q de Yule*. El cuadro *Transformar valores* de la Figura 14-11 permite estandarizar los valores de los datos, para los casos o las variables, antes de calcular las proximidades (no está disponible para datos binarios). Los métodos disponibles de estandarización (Figura 14-14) son: *Puntuaciones Z*, *Rango -1 a 1*, *Rango 0 a 1*, *Magnitud máxima de 1*, *Media de 1* y *Desviación típica 1*. El cuadro *Transformar medidas* de la Figura 14-11 permite transformar los valores generados por la medida de distancia. Las opciones disponibles son: *Valores absolutos*, *Cambiar el signo* y *Cambiar la escala al rango 0-1*.

El botón *Guardar* de la Figura 14-11 permite guardar información sobre la solución como nuevas variables para que puedan ser utilizadas en análisis subsiguientes. Estas variables son: *Conglomerado de pertenencia*, que permite guardar los conglomerados de pertenencia para una solución única o un rango de soluciones. Las variables guardadas pueden emplearse en análisis posteriores para explorar otras diferencias entre los grupos.

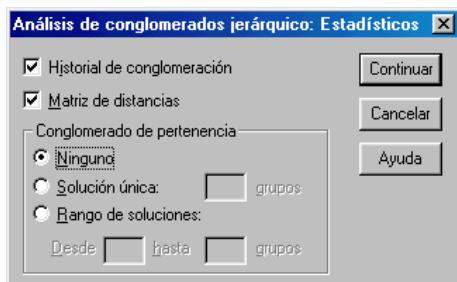


Figura 14-10

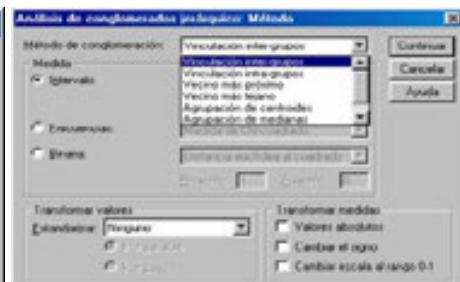


Figura 14-11

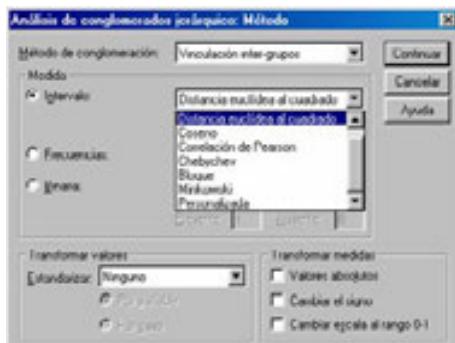


Figura 14-12

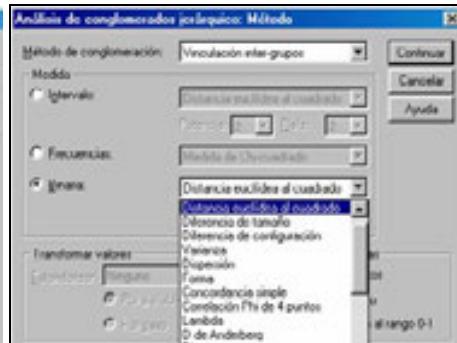


Figura 14-13

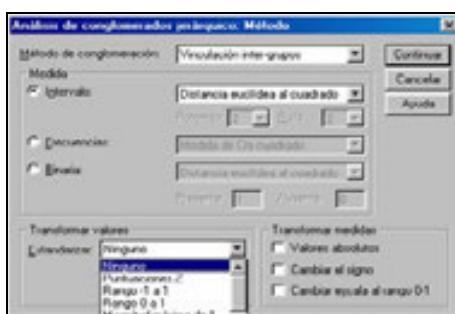


Figura 14-14

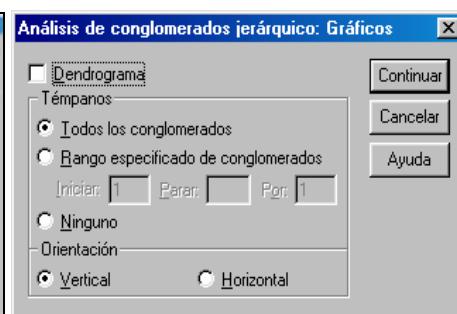


Figura 14-15

El botón *Gráficos* de la Figura 14-19 nos lleva a la pantalla de la Figura 14-15 cuya opción *Dendrograma* realiza el dendrograma correspondiente. Los dendrogramas pueden emplearse para evaluar la cohesión de los conglomerados que se han formado y proporcionar información sobre el número adecuado de conglomerados que deben conservarse. El dendrograma constituye la representación visual de los pasos de una solución de conglomeración jerárquica que muestra, para cada paso, los conglomerados que se combinan y los valores de los coeficientes de distancia. Las líneas verticales conectadas designan casos combinados. El dendrograma re-escala las distancias reales a valores entre 0 y 25, preservando la razón de las distancias entre los pasos. El cuadro *Témpanos* de la Figura 14-15 muestra un diagrama de témpanos, que incluye todos los conglomerados o un rango especificado de conglomerados. Los diagramas de témpanos muestran información sobre cómo se combinan los casos en los conglomerados, en cada iteración del análisis. La orientación permite seleccionar un diagrama vertical u horizontal: Diagrama de témpanos (Conglomerados). En la base de este diagrama (la derecha en los gráficos horizontales) no hay casos unidos todavía y a medida que se recorre hacia arriba el diagrama (o de derecha a izquierda en los horizontales), los casos que se unen se marcan con una X o una barra en la columna situada entre ellos, mientras que los conglomerados separados se indican con un espacio en blanco entre ellos.

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 14-9 para obtener los resultados del análisis cluster jerárquico según se muestra en la Figura 14-16. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. A continuación se presentan los centros iniciales de los conglomerados y el historial de iteraciones.

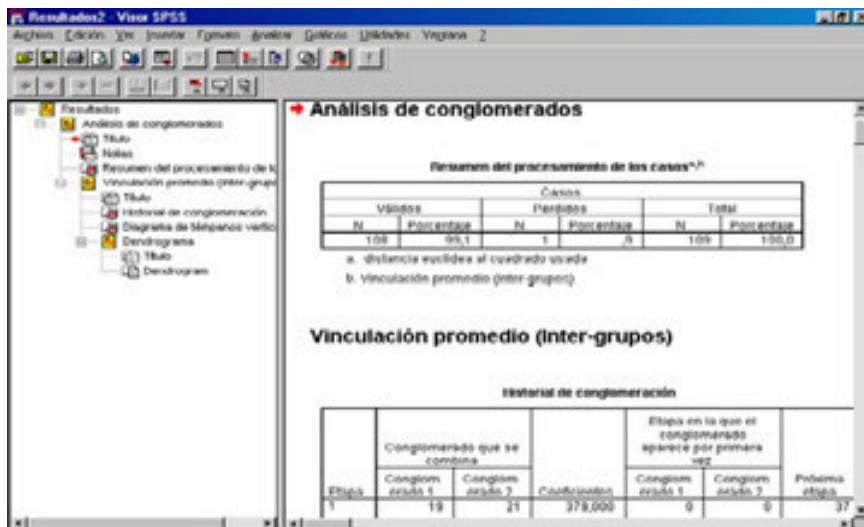


Figura 14-16

En la Figura 14-17 se presentan parte del dendograma correspondiente a este análisis cluster jerárquico.

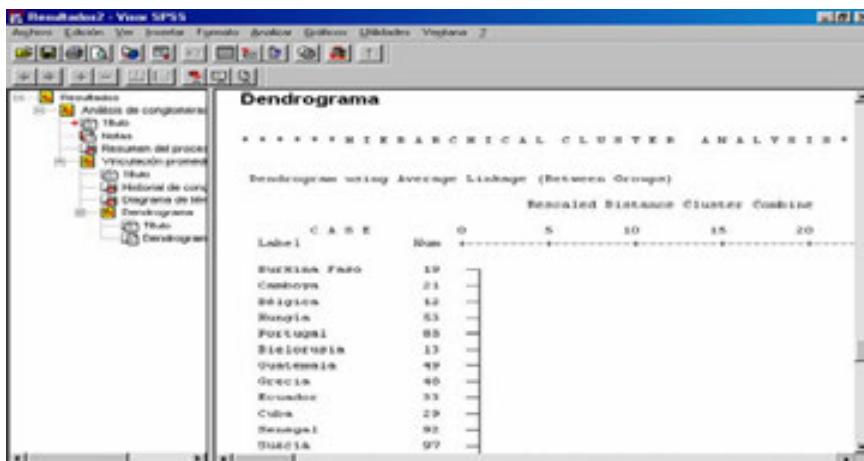


Figura 14-17

SPSS Y EL ANÁLISIS CLUSTER EN DOS FASES

El procedimiento *Análisis de conglomerados en dos fases* de SPSS es una herramienta de exploración diseñada para descubrir las agrupaciones naturales (o conglomerados) de un conjunto de datos que, de otra manera, no sería posible detectar. El algoritmo que emplea este procedimiento incluye varias atractivas funciones que lo hacen diferente de las técnicas de conglomeración tradicionales:

- *Tratamiento de variables categóricas y continuas*: Al suponer que las variables son independientes, es posible aplicar una distribución normal multinomial conjunta en las variables continuas y categóricas.
- *Selección automática del número de conglomerados*: Mediante la comparación de los valores de un criterio de selección del modelo para diferentes soluciones de conglomeración, el procedimiento puede determinar automáticamente el número óptimo de conglomerados.
- *Escalabilidad*: Mediante la construcción de un árbol de características de conglomerados (CF) que resume los registros, el algoritmo en dos fases puede analizar archivos de datos de gran tamaño.

Como ejemplo de aplicación, las empresas minoristas y de venta de productos para el consumidor suelen aplicar técnicas de conglomeración a los datos que describen los hábitos de consumo, sexo, edad, nivel de ingresos, etc. de los clientes. Estas empresas adaptan sus estrategias de desarrollo de productos y de marketing en función de cada grupo de consumidores para aumentar las ventas y el nivel de fidelidad a la marca.

Este procedimiento genera criterios de información (AIC o BIC) según el número de conglomerados de la solución, las frecuencias de los conglomerados para la conglomeración final y los estadísticos descriptivos por conglomerado para la conglomeración final. El procedimiento también genera gráficos de barras para las frecuencias de los conglomerados, gráficos de sectores para las frecuencias de los conglomerados y gráficos de la importancia de las variables. Además, proporciona medidas de la distancia que determinan cómo se calcula la similaridad entre dos conglomerados. Estas medidas son:

- *Log-verosimilitud*: La medida de la verosimilitud realiza una distribución de probabilidad entre las variables. Las variables continuas se supone que tienen una distribución normal, mientras que las variables categóricas se supone que son multinomiales. Se supone que todas las variables son independientes.

- *Euclídea*: La medida euclídea es la distancia según una "línea recta" entre dos conglomerados. Sólo se puede utilizar cuando todas las variables son continuas.

También existe una opción de *número de conglomerados* que permite especificar cómo se va a determinar el número de conglomerados. Hay dos formas:

- *Determinar automáticamente*: El procedimiento determinará automáticamente el número "óptimo" de conglomerados, utilizando el criterio especificado en el grupo Criterio de conglomeración. Si lo desea, introduzca un entero positivo para especificar el número máximo de conglomerados que el procedimiento debe tener en cuenta.
- *Especificar número fijo*: Permite fijar el número de conglomerados de la solución.

También existe una opción de *recuento de variables continuas* que proporciona un resumen de las especificaciones acerca de la tipificación de variables continuas realizadas en las opciones y una opción de *criterio de conglomeración* que determina cómo el algoritmo de conglomeración halla el número de conglomerados. Se puede especificar tanto el criterio de información bayesiano (BIC) como el criterio de información de Akaike (AIC).

Consideraciones previas

Este procedimiento trabaja tanto con variables continuas como categóricas. Los casos representan los objetos que se van a conglomerar y las variables representan los atributos en los que se va a basar la conglomeración. La medida de la distancia de la verosimilitud supone que las variables del modelo de conglomerados son independientes. Además, se supone que cada variable continua tiene una distribución normal (de Gauss) y que cada variable categórica tiene una distribución multinomial. Las comprobaciones empíricas internas indican que este procedimiento es bastante robusto frente a las violaciones tanto del supuesto de independencia como de las distribuciones, pero aún así es preciso tener en cuenta hasta qué punto se cumplen estos supuestos.

Por lo tanto, será conveniente utilizar el procedimiento *Correlaciones bivariadas* para comprobar la independencia de variables continuas y el procedimiento *Tablas de contingencia* para comprobar la independencia de dos variables categóricas. Utilice el procedimiento *Medias* para comprobar la independencia existente entre una variable continua y otra categórica. Utilice el procedimiento *Explorar* para comprobar la normalidad de una variable continua. Utilice el procedimiento *Prueba de chi-cuadrado* para comprobar si una variable categórica tiene una determinada distribución multinomial.

Ejercicio 14-1. Mediante análisis de conglomerados se trata de clasificar a los jóvenes (Fichero 14-1.sav) por el número de veces que van anualmente al fútbol, la paga semanal que reciben y el número de horas semanales que ven la televisión. Utilizar análisis cluster jerárquico y no jerárquico.

En el análisis cluster es necesario tipificar las variables, ya que, al trabajar con distancias, todas las variables han de venir medidas en las mismas unidades. Comenzamos entonces tipificando las variables afectadas (fútbol, paga2 y TV) rellenando la pantalla de entrada del procedimiento *Descriptivos* como se indica en la Figura 14-18 y su botón *Opciones* como se indica en la Figura 14-19. En la salida (Figura 14-20) se observa que la variación y el rango (según máximo y mínimo) de las tres variables son completamente distintos por lo que no hay comparabilidad posible de desviaciones típicas. Como en la Figura 14-18 se ha marcado la casilla *Guardar valores tipificados como variables*, al ejecutar el procedimiento se han obtenido tres nuevas variables tipificadas (zfútbol, zpaga2 y ztv).



Figura 14-18

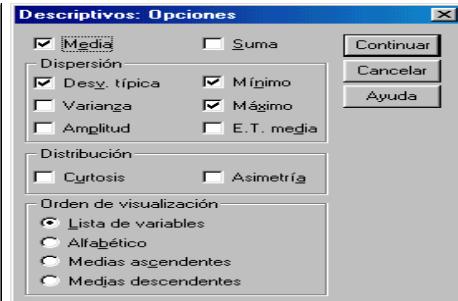


Figura 14-19

	N	Mínimo	Máximo	Media	Desv. típ.
ASISTENCIA ANUAL AL FUTBOL	14	0	8	3,71	3,43
PAGA SEMANAL EN PTAS	14	1000	2500	1557,14	730,35
HORAS SEMANALES TV	14	5	22	15,86	5,05
N válido (según lista)	14				

Figura 14-20

Si ahora volvemos a ejecutar el procedimiento *Descriptivos* con las variables tipificadas (Figura 14-21) se obtiene la salida de la Figura 14-22, que ya presenta rangos comparables para las tres variables.



Figura 14-21

	N	Mínimo	Máximo	Media	Desv. típ.
Puntu: ASISTENCIA ANUAL AL FUTBOL	14	-1,08319	1,24983	-4,2E-17	1,0000000
Puntu: PAGA SEMANAL EN PTAS	14	-,76285	1,29097	3,75E-16	1,0000000
Puntu: HORAS SEMANALES TV	14	-2,14934	1,21607	-9,0E-17	1,0000000
N válido (según lista)	14				

Figura 14-22

Otro paso interesante antes de realizar un análisis cluster es realizar un

gráfico de dispersión en tres dimensiones para las tres variables tipificadas con el objeto de atisbar los grupos que podrían formarse. Para ello elegimos *Gráficos* → *Dispersión*, seleccionamos 3-D (Figura 14-23) y rellenamos la pantalla de entrada del procedimiento *Diagramas de dispersión* como se indica en la Figura 14-24. Al pulsar *Aceptar* se obtiene el gráfico de dispersión para las variables tipificadas de la Figura 14-25, en el cual se intuye que podríamos agrupar a los individuos en tres conglomerados, ya que se observa una separación clara en tres grupos de puntos.

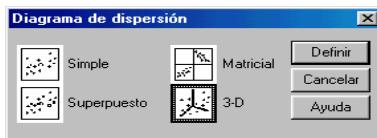


Figura 14-23

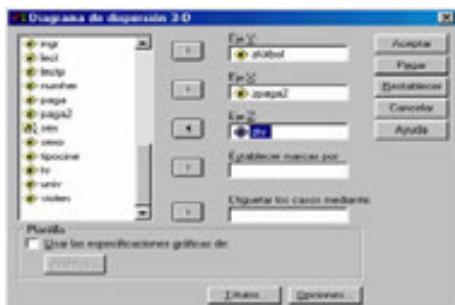


Figura 14-24

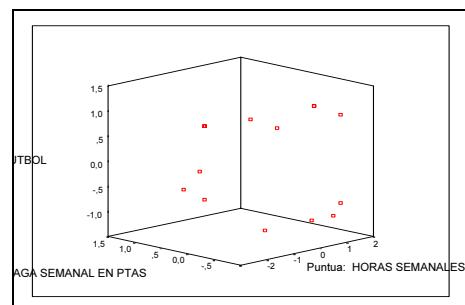


Figura 14-25

A continuación realizamos un análisis cluster jerárquico llenando la pantalla de entrada del procedimiento *Conglomerado de k-medias* como se indica en la Figura 14-26 y la pantalla de su botón *Opciones* según se indica en la Figura 14-27. La salida se muestra en las Figuras 14-28 a 14-30.

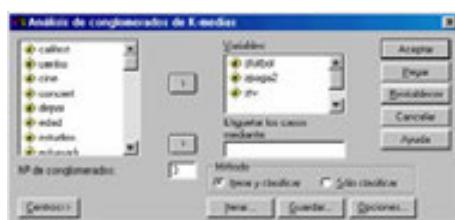


Figura 14-26

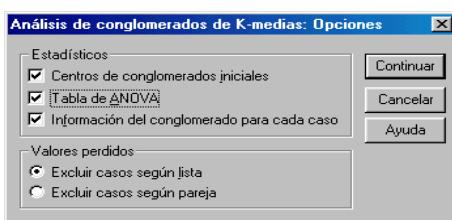


Figura 14-27

Análisis de conglomerados de K medias			
Centros iniciales de los conglomerados			
	Conglomerado		
	1	2	3
Puntu: ASISTENCIA ANUAL AL FUTBOL	1,24983	-,79156	,95821
Puntu: PAGA SEMANAL EN PTAS	-,76285	1,29097	-,76285
Puntu: HORAS SEMANALES TV	-2,14934	-,56562	1,21607

Historial de iteraciones*			
Iteración	Cambio en los centros de los conglomerados		
	1	2	3
1	,516	,753	,754
2	,000	,261	,243
3	,000	,000	,000

a. Covergencia alcanzada debido a un cambio en la distancia nulo o pequeño. La distancia máxima en la que ha cambiado cada centro es ,000. La iteración actual es 3. La distancia mínima entre los centros iniciales es 3,233.

Figura 14-28

Pertenencia de los conglomerados iniciales		
Número de caso	Conglomerado iniciales	Distancia
1	1	,516
2	2	,994
3	3	1,201
4	3	,700
5	2	,696
6	1	,516
7	3	,899
8	2	,584
9	3	1,281
10	3	1,266
11	2	,587
12	2	2,070
13	2	,591
14	3	1,216

Centros de los conglomerados finales			
	Conglomerado		
	1	2	3
Puntu: ASISTENCIA ANUAL AL FUTBOL	1,10492	-,45133	,00332
Puntu: PAGA SEMANAL EN PTAS	-,76285	,97149	-,71721
Puntu: HORAS SEMANALES TV	-1,05443	-,20268	,76415

Figura 14-29

Distancia entre los centros de los conglomerados iniciales			
Conglomerado	1	2	3
1		2,745	
2			2,816
3			2,013

ANOVA						
	Conglomerado		Error		F	Sig.
	Media cuadrática	gl	Media cuadrática	gl		
Puntu: ASISTENCIA ANUAL AL FUTBOL	1,851	2	,845	11	2,189	,158
Puntu: PAGA SEMANAL EN PTAS	4,956	2	,281	11	17,861	,000
Puntu: HORAS SEMANALES TV	4,567	2	,362	11	12,991	,001

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

Número de casos en cada conglomerado		
Conglomerado	1	2
1	2,000	
2		6,000
3		6,000
Válidos		14,000
Perdidos		,000

Figura 14-30

En la Figura 14-28 se presentan los centros iniciales de los conglomerados. Para el comienzo del método iterativo, en un principio se seleccionan tantos individuos como conglomerados hayamos solicitado de modo que estos individuos iniciales tengan distancia máxima entre ellos y que al estar separados lo suficiente produzcan los centros iniciales. Una vez estimados los centroides iniciales se calcula la distancia de cada punto a cada uno de ellos y en función de la mínima distancia obtenida se irán clasificando los individuos en los tres grupos de conglomerados. Realizados los tres grupos, se calculan los tres centros y se repite el mismo proceso para hacer otra agrupación, y así sucesivamente hasta agotar las iteraciones o hasta que se cumpla el criterio de parada. En el historial de iteraciones de la Figura 14-28 aparece el número de iteraciones realizadas y los cambios producidos en los centroides. En la Figura 14-29 se presentan los centros de los conglomerados obtenidos al final del proceso iterativo y la lista de pertenencia de cada individuo a su conglomerado con la distancia de cada uno al centro de su grupo.

En la Figura 14-30 se presenta una tabla ANOVA para los conglomerados cuyas pruebas F sólo se deben utilizar con una finalidad descriptiva, puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales. Lo relevante son los valores F, que no deben ser muy pequeños (lo más alejados posible del valor 1) para que las variables sean realmente efectivas en la identificación de clusters.

La tabla de pertenencia a los conglomerados de la Figura 14-29 permite realizar los siguientes clusters o conglomerados {1,6}, {2, 5, 8, 11, 12, 13} y {3, 4, 7, 9, 10, 14}.

Para realizar un análisis jerárquico de conglomerados, rellenamos la pantalla de entrada del procedimiento *Conglomerados jerárquicos* como se indica en la Figura 14-31 y las pantallas Gráficos, Estadísticos y Método según las Figuras 14-32 a 14-34. La salida se muestra en las Figuras 14-35 a 14-37. El dendograma sugiere los conglomerados {3,9,4}, {7,10,14}, {2,8,5,11,13} y {1,6,12}, que no están muy lejos de los del caso anterior (si unimos los dos primeros).

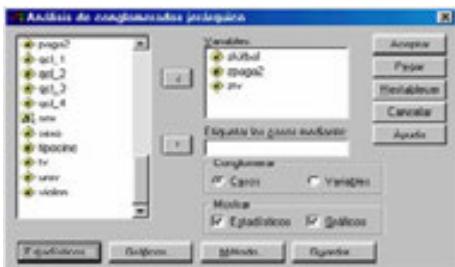


Figura 14-31



Figura 14-32

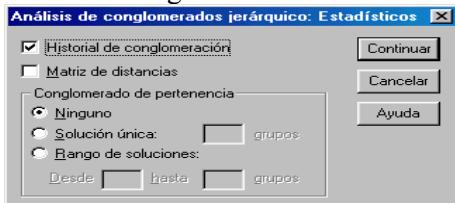


Figura 14-33



Figura 14-34

Vinculación promedio (Inter-grupos)									
Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa			
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2				
	1	3	,000	0	0	8			
2		2	,000	0	0	9			
3		7	,104	0	0	6			
4		5	,379	0	0	5			
5		5	,575	4	0	9			
6		7	,679	3	0	10			
7		1	,1,065	0	0	11			
8		3	,1,065	1	0	10			
9		2	,1,640	2	5	12			
10		3	,5,138	8	6	12			
11		1	,5,157	7	0	13			
12		2	,6,565	9	10	13			
13		1	,8,378	11	12	0			

Figura 14-35

Diagrama de témpanos vertical

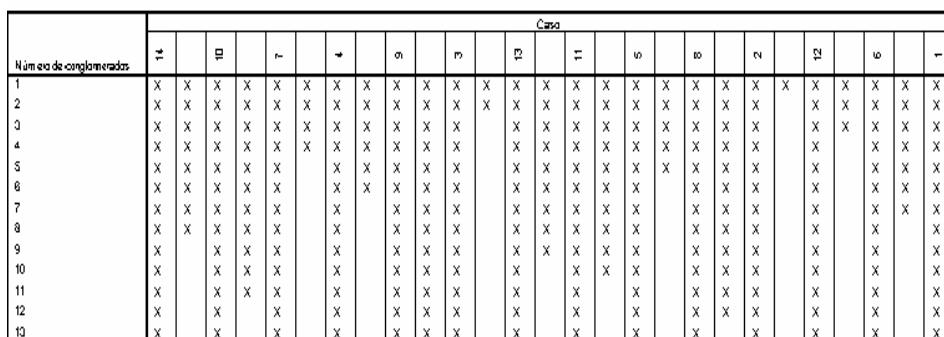


Figura 14-36

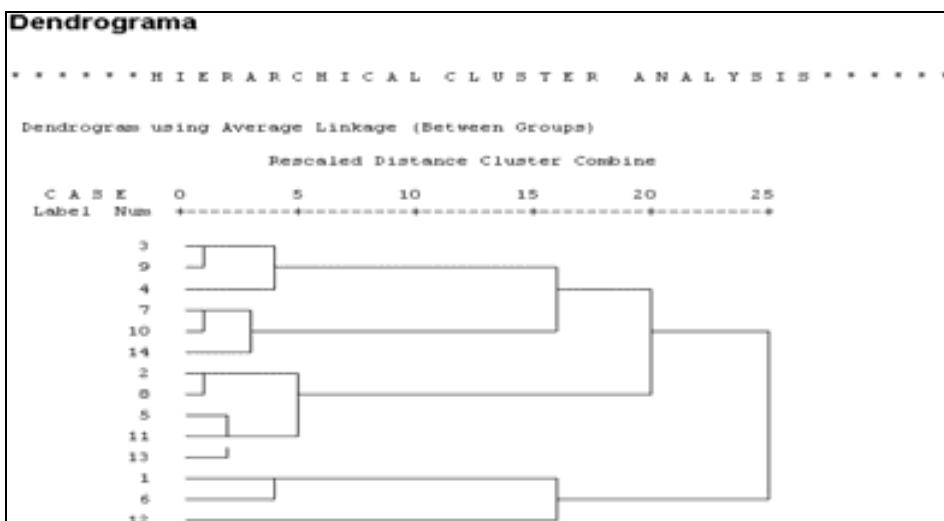


Figura 14-37

Ejercicio 14-2. Mediante análisis de conglomerados se trata de clasificar automóviles (Fichero 14-2.sav) de acuerdo a sus precios y a sus propiedades físicas. Utilizar análisis cluster en dos fases.

Comenzamos eligiendo *Analizar* → *Clasificar* → *Conglomerados en dos fases* (figura 14-38) y rellenando la pantalla *Análisis de conglomerados en dos fases* como se indica en la figura 14-39. Como variable categórica se utiliza el tipo de vehículo (*Vehicle type*) y como variables continuas se usan desde *Price in thousands* hasta *Fuel efficiency*. Hacemos clic en *Gráficos*, rellenamos la pantalla como se indica en la figura 14-40 y pulsamos *Continuar*. Hacemos clic en *Resultados*, rellenamos la pantalla como se indica en la figura 14-41 y pulsamos *Continuar*. Hacemos clic en *Opciones* para fijar el tratamiento de valores atípicos, la tipificación de variables y al asignación de memoria (figura 14-42). pulsamos *Continuar* y *Aceptar* y ya obtenemos las salida del análisis cluster en dis fases.

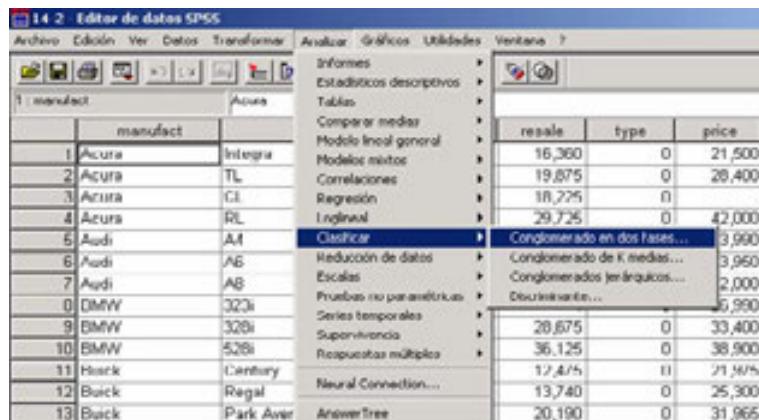


Figura 14-38

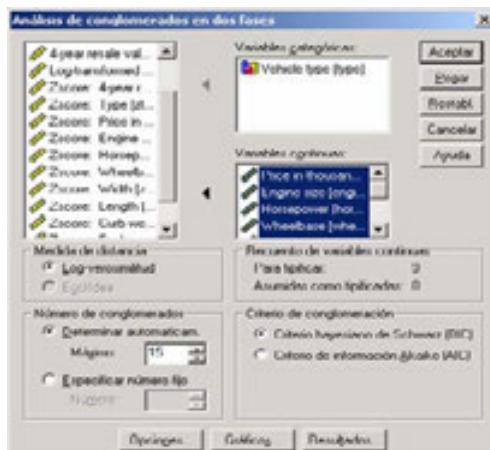


Figura 14-39

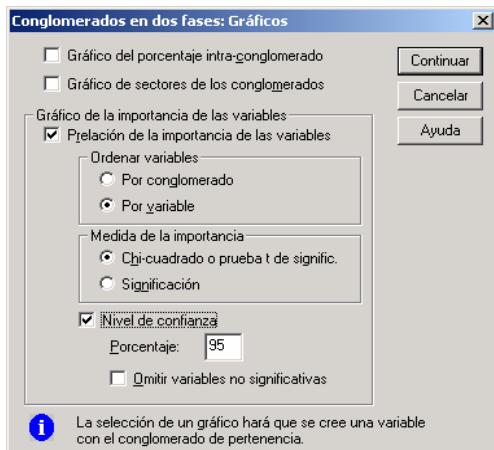


Figura 14-40

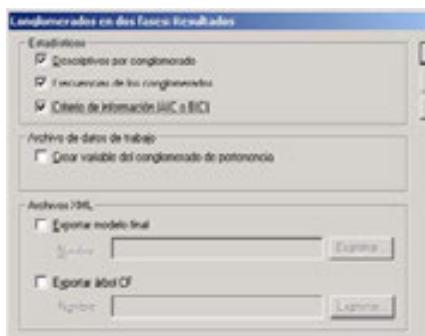


Figura 14-41

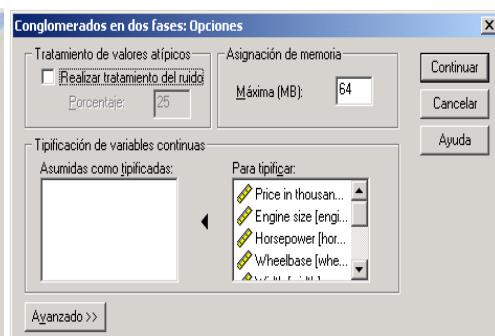


Figura 14-42

La primera parte de la salida es un informe sobre las posibles agrupaciones en conglomerados. Inicialmente el número de conglomerados adecuado es aquél que tiene un mayor BIC, pero como hay tramos de BIC decreciente creciendo el número de conglomerados, será necesario considerar la tasa de cambio (no unitaria) del BIC simultáneamente con el propio BIC y eligiendo como número de conglomerados el correspondiente a los mayores de BIC y su tasa de cambio simultáneamente. Por tanto, se formarán tres conglomerados, cuya distribución de observaciones se muestra en la parte siguiente de la salida.

Conglomerados en dos fases

Agrupación automática

Número de conglomerados	Criterio bayesiano de Schwarz (BIC)	Cambio en BIC(a)	Razón de cambios en BIC(b)	Razón de medidas de distancia(c)
1	1214,377			
2	974,051	-240,326	1,000	1,829
3	885,924	-88,128	,367	2,190
4	897,559	11,635	-,048	1,368
5	931,760	34,201	-,142	1,036
6	968,073	36,313	-,151	1,576
7	1026,000	57,927	-,241	1,083
8	1086,815	60,815	-,253	1,687
9	1161,740	74,926	-,312	1,020
10	1237,063	75,323	-,313	1,239
11	1316,271	79,207	-,330	1,046
12	1396,192	79,921	-,333	1,075
13	1477,199	81,008	-,337	1,076
14	1559,230	82,030	-,341	1,301
15	1644,366	85,136	-,354	1,044

a Los cambios proceden del número anterior de conglomerados de la tabla.

b Las razones de los cambios están relacionadas con el cambio para la solución de los dos conglomerados.

c Las razones de las medidas de la distancia se basan en el número actual de conglomerados frente al número de conglomerados anterior.

Distribución de conglomerados

		N	% de combinados	% del total
Conglomerado	1	62	40,8%	39,5%
	2	39	25,7%	24,8%
	3	51	33,6%	32,5%
	Combinados	152	100,0%	96,8%
Casos excluidos		5		3,2%
Total		157		100,0%

Se observa que de los 157 casos totales, 5 se excluyeron del análisis debido al efecto de los valores perdidos. De los 152 casos asignados a los clusters, 62 (40,8%) se asignaron al primer cluster, 39 al segundo (25,7%) y 51 al tercero (33,6%). La última columna presenta los porcentajes respecto al número total de casos (sin desaparecidos).

La tabla de frecuencias por tipo de vehículo (automóviles o camiones) clarifica las propiedades de los clusters según los valores de la variable cualitativa considerada. Por ejemplo, el segundo cluster está formado exclusivamente por camiones y el tercero exclusivamente por automóviles, mientras que el primero tiene un alto porcentaje de automóviles y un sólo coche (2,5% del total).

Vehicle type

		Automobile		Truck	
		Frecuencia	Porcentaje	Frecuencia	Porcentaje
Conglomerado	1	61	54,5%	1	2,5%
	2	0	,0%	39	97,5%
	3	51	45,5%	0	,0%
	Combinados	112	100,0%	40	100,0%

Los gráficos por variables producen un gráfico separado por cada cluster. Las variables se sitúan en el eje de ordenadas con valores decrecientes en cuanto a su importancia en la formación de los clusters. Las líneas verticales con guiones muestran los valores críticos para determinar la significatividad de cada variable en la formación del cluster. Una variable es significativa si el estadístico T de Student excede la línea de guiones positiva o negativa. Las variables que resulten significativas contribuyen a la formación del cluster. Un valor negativo de la T indica que la variable toma valor en el cluster inferior a su media y un valor positivo indica lo contrario. Para el cluster 1 la variable *Fuell efficiency* toma valores mayores que su valor medio (figura 14-43) y el resto de las variables toma valores menores y todas las variables tiene importancia en la formación del cluster. Para el cluster 2 ocurre el complementario (figura 14-44) salvo que las variables *Width*, *Length*, *Horsepower* y *Price in thousands* no tienen importancia en al formación del cluster porque no alcanzan la línea discontinua de la T. Para el tercer conglomerado se mantiene la misma tónica, pero las variables que no alcanzan la línea de guiones son *Whelbase* y *Fuell capacity* (figura 14-45).

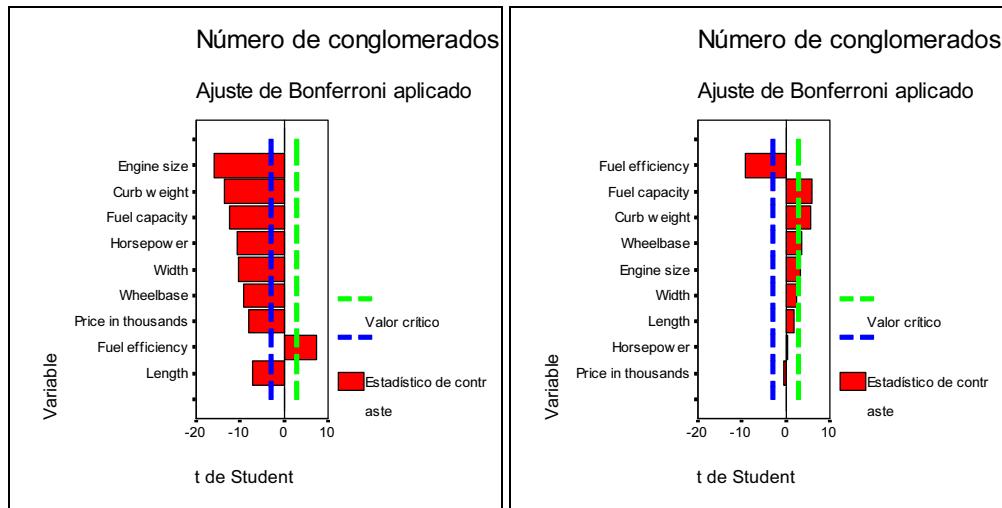


Figura 14-43

Figura 14-43

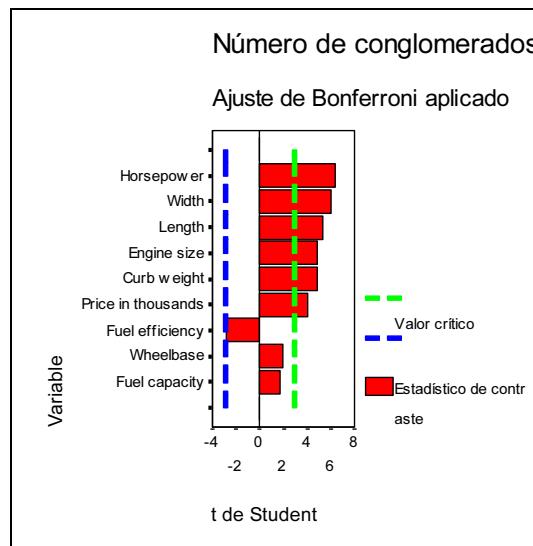


Figura 14-45

También se obtienen intervalos de confianza al 95% para las medias de las variables cuantitativas en los tres conglomerados, divididos por una línea que indica la pertenencia o no a cada una de las dos clases de la variable categórica. Las figuras 14-46 a 14-50 representan estos intervalos de confianza para algunas de las variables cuantitativas consideradas.

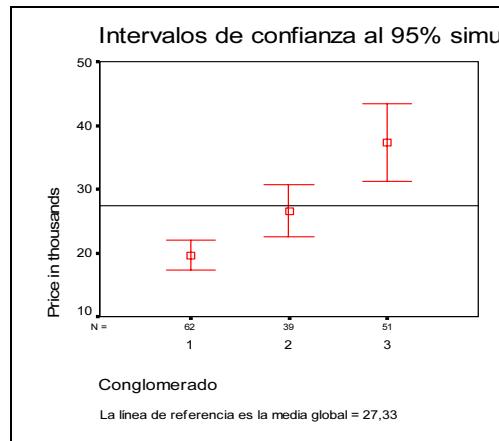


Figura 14-46

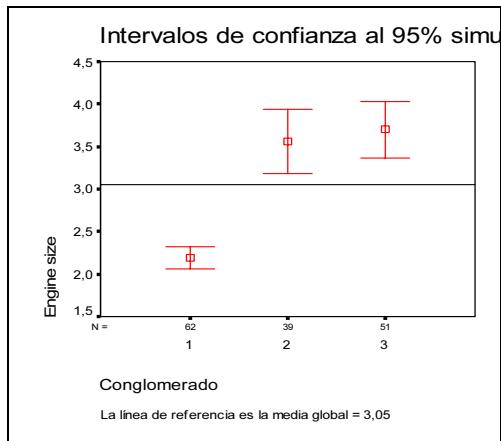


Figura 14-47

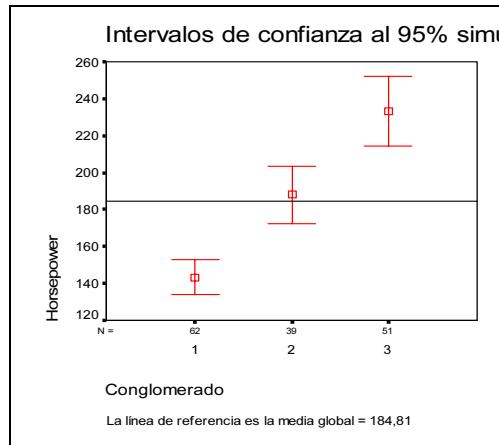


Figura 14-48

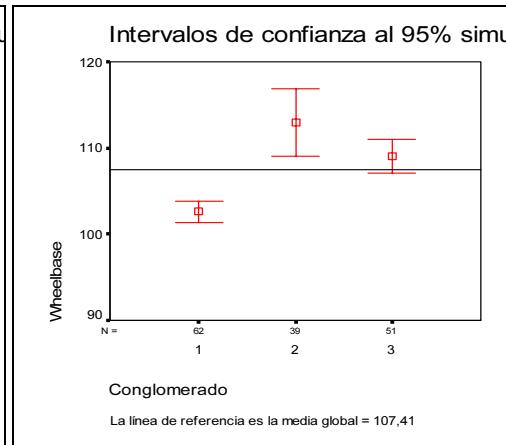


Figura 14-49

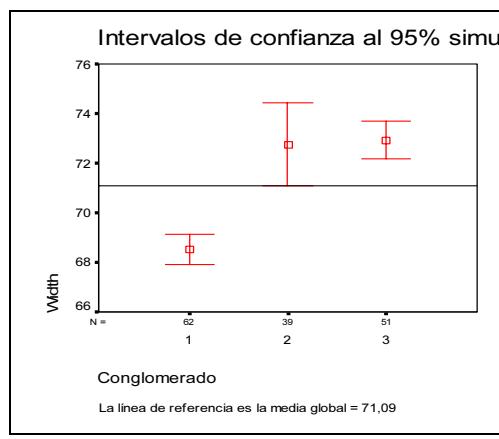


Figura 14-50

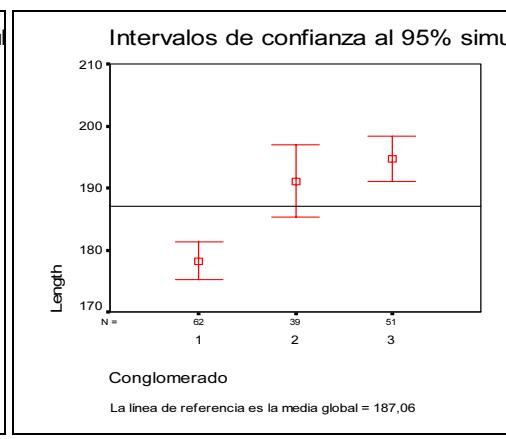


Figura 14-51

CLASIFICACIÓN Y SEGMENTACIÓN MEDIANTE ANÁLISIS DISCRIMINANTE

CONCEPTO DE ANÁLISIS DISCRIMINANTE

El análisis discriminante es una técnica estadística que permite asignar o clasificar nuevos individuos dentro de grupos previamente reconocidos o definidos. Para ilustrar el concepto, consideremos un ejemplo típico en el campo de la medicina. Supongamos que se dispone de una muestra de pacientes en los que se ha medido un conjunto de variables relativas al diagnóstico de una enfermedad (presión sanguínea, edad, peso, etc.) y que con esta información o por comprobación posterior, el investigador ha dividido la muestra en dos (o más) grupos diagnósticos. La finalidad del análisis discriminante es que cuando llegue un nuevo enfermo en el que son medidas las mismas variables, sus valores permitan asignar dicho paciente a un grupo de diagnóstico con la máxima probabilidad, cuantificando a la vez el valor de esta probabilidad. El análisis discriminante puede aplicarse a todos los campos de la ciencia en los que el objeto de investigación sea la clasificación de individuos, a través de un perfil observado. El análisis discriminante se conoce en ocasiones como *análisis de la clasificación*, ya que su objetivo fundamental es producir una regla o un esquema de clasificación que permita a un investigador predecir la población a la que es más probable que tenga que pertenecer una nueva observación (supuestas conocidas varias poblaciones a las que pueden pertenecer las observaciones).

El análisis parte de una tabla de datos de n individuos en que se han medido p variables cuantitativas independientes o “explicativas”, como perfil de cada uno de ellos. Una variable cualitativa adicional (dependiente o “clasificativa”), con dos (o más) categorías, ha definido por otros medios el grupo a que cada individuo pertenece. A partir de esta variable cualitativa se obtendrá un modelo matemático discriminante contra el cual será contrastado el perfil de un nuevo individuo cuyo grupo se desconoce para, en función de un resultado numérico, ser asignado al grupo más probable. Cuanto mejor sea la información de partida más fiable será el resultado de asignaciones posteriores.

El análisis discriminante persigue *explicar* la pertenencia de cada individuo original a uno u otro grupo preestablecido, en función de las variables de su perfil, y a la vez que cuantificar el peso de cada una de ellas en la discriminación. Por otro lado, el análisis discriminante persigue *predecir* a qué grupo más probable habrá de pertenecer un nuevo individuo del que únicamente se conoce su perfil de variables. La variable categórica “grupo” es lo que se explica y lo que predice.

En la clasificación discriminante hay dos enfoques. El primero de ellos está basado en la obtención de funciones discriminantes de cálculo similar a las ecuaciones de regresión lineal múltiple. El segundo emplea técnicas de correlación canónica y de componentes principales y se denomina *análisis discriminante canónico*. El primero es el más común y su fundamento matemático está en conseguir, a partir de las variables explicativas, unas funciones lineales de éstas con capacidad para clasificar otros individuos. A cada nuevo caso se aplican dichas ecuaciones, y la función de mayor valor define el grupo a que pertenece.

CLASIFICACIÓN CON DOS GRUPOS

Se trata de estudiar la aplicación del análisis discriminante a la clasificación de individuos en el caso de que dichos individuos se puedan asignar solamente a dos grupos a partir de k variables clasificadoras. Este problema lo resolvió Fisher analíticamente mediante su función discriminante.

La *función discriminante de Fisher D* se obtiene como función lineal de k variables explicativas como:

$$D = u_1 X_1 + u_2 X_2 + \cdots + u_k X_k$$

Se trata de obtener los coeficientes de ponderación u_j . Si consideramos que existen n observaciones, podemos expresar la función discriminante para ellas:

$$D_i = u_1 X_{1i} + u_2 X_{2i} + \cdots + u_k X_{ki} \quad i = 1, 2, \dots, n$$

D_i es la puntuación discriminante correspondiente a la observación i -ésima. Expresando las variables explicativas en desviaciones respecto a la media, D_i también lo estará y la relación anterior se puede expresar en forma matricial como sigue:

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$$

En notación compacta podemos escribir:

$$\mathbf{d} = \mathbf{X}\mathbf{u}$$

La variabilidad de la función discriminante (suma de cuadrados de las variables discriminantes en desviaciones respecto a su media) se expresa como:

$$\mathbf{d}'\mathbf{d} = \mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u}$$

La matriz $\mathbf{X}'\mathbf{X}$ es una matriz simétrica expresada en desviaciones respecto a la media, por lo que puede considerarse como la matriz \mathbf{T} de suma de cuadrados (SCPC) total de las variables (explicativas) de la matriz \mathbf{X} . Según la teoría del análisis multivariante de la varianza, $\mathbf{X}'\mathbf{X}$ se puede descomponer en la suma de la matriz entre grupos \mathbf{F} y la matriz intragrupo \mathbf{V} (o residual). Se tiene:

$$\mathbf{X}'\mathbf{X} = \mathbf{T} = \mathbf{F} + \mathbf{V}$$

Por lo tanto:

$$\mathbf{d}'\mathbf{d} = \mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} = \mathbf{u}'\mathbf{T}\mathbf{u} = \mathbf{u}'\mathbf{F}\mathbf{u} + \mathbf{u}'\mathbf{V}\mathbf{u}$$

En la igualdad anterior \mathbf{T} , \mathbf{F} y \mathbf{V} son calculables con los datos muestrales mientras que los coeficientes u_i están por determinar. Fisher obtuvo los u_i maximizando la razón de la variabilidad entre grupos respecto de la variabilidad intragrupo. La razón de ser de este criterio es la obtención del eje discriminante de forma que las distribuciones proyectadas sobre el mismo estén lo más separadas posible entre sí (mayor variabilidad entre grupos) y, al mismo tiempo, que cada una de las distribuciones esté lo menos dispersa (menor variabilidad dentro de los grupos). Analíticamente, el criterio de Fisher nos lleva a la maximización de λ , donde:

$$\lambda = \frac{\mathbf{u}'\mathbf{F}\mathbf{u}}{\mathbf{u}'\mathbf{W}\mathbf{u}}$$

La solución a este problema se obtiene derivando λ respecto de \mathbf{u} e igualando a cero, es decir:

$$\frac{\partial \lambda}{\partial \mathbf{u}} = \frac{2\mathbf{F}\mathbf{u}(\mathbf{u}'\mathbf{W}\mathbf{u}) - 2\mathbf{W}\mathbf{u}(\mathbf{u}'\mathbf{F}\mathbf{u})}{(\mathbf{u}'\mathbf{W}\mathbf{u})^2} = 0 \Rightarrow 2\mathbf{F}\mathbf{u}(\mathbf{u}'\mathbf{W}\mathbf{u}) - 2\mathbf{W}\mathbf{u}(\mathbf{u}'\mathbf{F}\mathbf{u}) = 0$$

De donde:

$$\frac{2Fu}{2Wu} = \frac{u'Fu}{u'Wu} = \lambda \Rightarrow Fu = Wu\lambda \Rightarrow W^{-1}Fu = \lambda u$$

Por lo tanto, la ecuación para la obtención del primer eje discriminante $W^{-1}Fu = \lambda u$ se traduce en la obtención de un vector propio u asociado a la matriz no simétrica $W^{-1}F$.

De los valores propios λ que se obtienen al resolver la ecuación $W^{-1}Fu = \lambda u$ se retiene el mayor, ya que precisamente λ es la ratio que queremos maximizar y u es el vector característico asociado al mayor valor propio de la matriz $W^{-1}F$.

Dado que λ es la ratio a maximizar, nos medirá, una vez calculado, el poder discriminante del primer eje discriminante. En nuestro caso no necesitamos más ejes discriminantes, pues estamos haciendo análisis discriminante con dos grupos.

En un caso general de análisis discriminante con G grupos ($G > 2$), el número máximo de ejes discriminantes que se pueden obtener viene dado por $\min(G-1, k)$. Por lo tanto pueden obtenerse hasta $G-1$ ejes discriminantes, si el número de variables explicativas k es mayor o igual que $G-1$, hecho que suele ser siempre cierto, ya que en las aplicaciones prácticas el número de variables explicativas suele ser grande.

El resto de los ejes discriminantes vendrán dados por los vectores propios asociados a los valores propios de la matriz $W^{-1}F$ ordenados de mayor a menor. Así, el segundo eje discriminante tendrá menos poder discriminatorio que el primero, pero más que cualquiera de los restantes.

Como la matriz $W^{-1}F$ no es simétrica, los ejes discriminantes no serán en general ortogonales (perpendiculares entre sí).

En el caso de análisis discriminante con dos grupos, los coeficientes u_1, u_2, \dots, u_k normalizados correspondientes a las coordenadas del vector propio unitario asociado al mayor valor propio de la matriz $W^{-1}F$ obtenidos en el proceso de maximización, pueden contemplarse como un conjunto de cosenos directores que definen la situación del eje discriminante.

Las **puntuaciones discriminantes** son pues los valores que se obtienen al dar valores a X_1, X_2, \dots, X_k en la ecuación:

$$D = u_1 X_1 + u_2 X_2 + \cdots + u_k X_k$$

Las puntuaciones discriminantes se corresponden con los valores obtenidos al proyectar cada punto del espacio k-dimensional de las variables originales sobre el eje discriminante.

Los **centros de gravedad o centroides** (vector de medias) son los estadísticos básicos que resumen la información sobre los grupos. Los centroides de los grupos *I* y *II* serán los siguientes:

$$\bar{x}_I = \begin{bmatrix} \bar{X}_{1,I} \\ \bar{X}_{2,I} \\ \vdots \\ \bar{X}_{k,I} \end{bmatrix} \quad \bar{x}_{II} = \begin{bmatrix} \bar{X}_{1,II} \\ \bar{X}_{2,II} \\ \vdots \\ \bar{X}_{k,II} \end{bmatrix}$$

Con lo que, para los grupos I y II se obtiene:

$$\bar{D}_I = u_1 \bar{X}_{1,I} + u_2 \bar{X}_{2,I} + \cdots + u_k \bar{X}_{k,I}$$

$$\bar{D}_{II} = u_1 \bar{X}_{1,II} + u_2 \bar{X}_{2,II} + \cdots + u_k \bar{X}_{k,II}$$

El **punto de corte discriminante** *C* se calcula mediante el promedio:

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2}$$

El criterio para clasificar el individuo *i* es el siguiente:

- Si $D_i < C$, se clasifica al individuo *i* en el grupo *I*
- Si $D_i > C$, se clasifica al individuo *i* en el grupo *II*

En general, cuando se aplica el análisis discriminante se le resta el valor de *C* a la función discriminante, que vendrá dada por:

$$D - C = u_1 X_1 + u_2 X_2 + \cdots + u_k X_k - C$$

En este último caso, se clasifica a un individuo en el grupo *I* si $D - C > 0$, y en el grupo *II* en otro caso.

A veces suelen construirse funciones discriminantes para cada grupo, F_I y F_{II} , con la siguiente estructura:

$$F_I = a_{I,1}X_1 + a_{I,2}X_2 + \cdots + a_{I,k}X_k - C_I$$

$$F_{II} = a_{II,1}X_1 + a_{II,2}X_2 + \cdots + a_{II,k}X_k - C_{II}$$

Cuando se utilizan estas funciones, se clasifica un individuo en el grupo para el que la función F_j sea mayor. Este tipo de funciones clasificadoras tienen la ventaja de que se generalizan fácilmente al caso de que existan más de dos grupos y vienen implementadas en la mayoría del software estadístico.

Si hacemos:

$$\begin{aligned} F_{II} - F_I &= (a_{II,1} - a_{I,1})X_1 + (a_{II,2} - a_{I,2})X_2 + \cdots + (a_{II,k} - a_{I,k})X_k - (C_{II} - C_I) \\ &= u_1X_1 + u_2X_2 + \cdots + u_kX_k - C = D - C \end{aligned}$$

ya se pueden obtener los coeficientes u_1, u_2, \dots, u_k .

Existen otros criterios de clasificación, entre los que destacan el análisis de la regresión y la distancia de Mahalanobis.

La **relación entre el análisis de la regresión y el análisis discriminante** con dos grupos es muy estrecha. Si se realiza un ajuste por mínimos cuadrados, tomando como variable dependiente la variable dependiente que define la pertenencia a uno u otro grupo y como variables explicativas a las variables clasificadoras, se obtienen unos coeficientes que guardan una estricta proporcionalidad con la función discriminante de Fisher.

La **distancia de Mahalanobis** es una generalización de la distancia euclídea que tiene en cuenta la matriz de covarianzas intragrupos. El cuadrado de la distancia de Mahalanobis \mathbf{DM}_{ij}^2 entre los puntos i y j en un espacio de p dimensiones, siendo \mathbf{V}_w la matriz de covarianzas intragrupos, viene definida por:

$$\mathbf{DM}_{i,j}^2 = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{V}_w^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

donde los vectores \mathbf{x}_i y \mathbf{x}_j representan dos puntos en el espacio p-dimensional.

La distancia euclídea es un caso particular de la distancia de Mahalanobis en la que $V_w = \mathbf{I}$. La distancia euclídea no tiene en cuenta la dispersión de las variables y las relaciones existentes entre ellas, mientras que en la distancia de Mahalanobis sí que se descuentan estos factores al introducir la inversa de la matriz de covarianzas intragrupo. La distancia euclídea será:

$$d_{i,j}^2 = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{I} (\mathbf{x}_i - \mathbf{x}_j) = \sum_{h=1}^p (X_{ih} - X_{jh})^2$$

Con el criterio de la distancia de Mahalanobis se calculan, para el punto i , las dos distancias siguientes:

$$DM_{i,I}^2 = (\mathbf{x}_i - \mathbf{x}_I)' V_w^{-1} (\mathbf{x}_i - \mathbf{x}_I)$$

$$DM_{i,II}^2 = (\mathbf{x}_i - \mathbf{x}_{II})' V_w^{-1} (\mathbf{x}_i - \mathbf{x}_{II})$$

La aplicación de este criterio consiste en asignar cada individuo al grupo para el que la distancia de Mahalanobis es menor.

Se observa que la distancia de Mahalanobis se calcula en el espacio de las variables originales, mientras que en el criterio de Fisher se sintetizan todas las variables en la función discriminante, que es la utilizada para realizar la clasificación.

CONTRASTES Y PROBABILIDAD DE PERTENENCIA (2 GRUPOS)

Realizando determinadas hipótesis se pueden ejecutar contrastes de significación sobre el modelo discriminante así como contrastes para seleccionar las variables cuando el número de éstas es muy grande y no se conoce a priori las variables que son relevantes en el análisis.

Por otra parte, el cálculo de probabilidad de pertenencia a un grupo requiere que previamente se haya postulado algún modelo probabilístico de la población.

Las hipótesis estadísticas que se adoptan, análogas a las postuladas en el análisis multivariante de la varianza, se refieren tanto a la población como al proceso de obtención de la muestra. Las hipótesis sobre la población son las siguientes:

- *Hipótesis de homoscedasticidad:* La matriz de covarianzas de todos los grupos es constante igual a Σ .

- *Hipótesis de normalidad:* Cada uno de los grupos tiene una distribución normal multivariante, es decir, $x_g \rightarrow N(\mu_g, \Sigma)$

La hipótesis sobre el proceso de obtención de la muestra facilita la realización del proceso de inferencia a partir de la información disponible. Dicha hipótesis consiste en suponer que se ha extraído una muestra aleatoria multivariante independiente en cada uno de los G grupos.

Bajo las hipótesis anteriores, la función discriminante obtenida por Fisher es óptima. No obstante, la hipótesis de que las variables clasificadoras sigan una distribución normal no sería razonable para variables categóricas, utilizadas frecuentemente en el análisis discriminante como variables clasificadoras. Conviene señalar que, cuando se utilizan variables de este tipo, la función discriminante lineal de Fisher no tiene el carácter de óptima. A continuación, y basados en las hipótesis anteriores, se examinan los contrastes de significación del modelo, el problema de selección de variables y el cálculo de probabilidades de pertenencia a una población.

Con los *contrastos de significación y evaluación de la bondad del ajuste* que se realizan en el análisis discriminante con dos grupos, se trata de dar respuesta a tres tipos de cuestiones diferentes:

- ¿Se cumple la hipótesis de homoscedasticidad del modelo?
- ¿Se cumple la hipótesis de normalidad?
- ¿Difieren significativamente las medias poblacionales de los dos grupos?

La justificación de las primeras cuestiones ya se conoce de la teoría de modelos. El análisis de normalidad en el caso multivariante se suele realizar variable a variable, dada la complejidad de hacerlo conjuntamente. Para el contraste de homoscedasticidad se puede utilizar el estadístico de Barlett-Box.

La respuesta que se dé a la cuestión c) es crucial para la justificación de la realización del análisis discriminante. En el caso de que la respuesta fuese negativa carecería de interés continuar con el análisis, ya que significaría que las variables introducidas como variables clasificadoras no tienen una capacidad discriminante *significativa*. La hipótesis nula y alternativa para dar respuesta a la cuestión c) son las siguientes: $H_0 : \mu_1 = \mu_2$ y $H_1 : \mu_1 \neq \mu_2$.

El contraste de la hipótesis anterior se puede realizar específicamente mediante el estadístico T^2 de Hotelling definido como sigue:

$$\mathbf{T}^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \bar{\mathbf{S}}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \left(\frac{\mathbf{n}_1 \mathbf{n}_2}{\mathbf{n}_1 + \mathbf{n}_2} \right)$$

donde:

$$\bar{\mathbf{S}} = \frac{\mathbf{W}_1 + \mathbf{W}_2}{\mathbf{n}_1 + \mathbf{n}_2 - 2}$$

La matriz $\bar{\mathbf{S}}$ es un estimador insesgado de la matriz de covarianzas poblacional Σ , obtenido bajo el supuesto de que la matriz de covarianzas poblacional es la misma en los dos grupos. \mathbf{W}_1 y \mathbf{W}_2 son las sumas de cuadrados y productos cruzados calculados para cada grupo ($\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2$).

Bajo la hipótesis nula ($H_0 : \mu_1 = \mu_2$), la T^2 de Hotelling tiene una distribución relacionada con la F de Fisher Snedecor como sigue:

$$\frac{n_1 + n_2 - k - 1}{k} \frac{T^2}{n_1 + n_2 - 2} \rightarrow F_{k, n_1 + n_2 - k - 1}$$

Existen otros estadísticos que se pueden emplear, diseñados para el caso general de G grupos, tales como el estadístico Ra de Rao o el estadístico V de Barlett que están construidos a partir de la Λ de Wilks como sigue:

$$Ra = \frac{1 - \Lambda^{\frac{1}{s}}}{\Lambda^{\frac{1}{s}}} \frac{1 + ts - k(G-1)/2}{k(G-1)} \rightarrow F_{k(G-1), 1+ts-k(G-1)/2}$$

$$t = n - 1 - (k + G)/2 \quad s = \sqrt{\frac{k^2(G-1)^2 - 4}{k^2 + (G-1)^2 - 5}}$$

$$V = -\left\{ n - 1 - \frac{k + G}{2} \right\} \ln(\Lambda) \rightarrow \chi^2_{k(G-1)} \quad \Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

En caso de que se rechace la hipótesis nula ($H_0 : \mu_1 = \mu_2$) se puede aplicar el análisis univariante de la varianza para contrastar la hipótesis de igualdad de medias para cada una de las variables clasificadoras por separado.

Como medida de evaluación de la bondad del ajuste se utiliza el coeficiente *eta cuadrado* (η^2), que es el coeficiente de determinación obtenido al realizar la regresión entre la variable dicotómica, que indica la pertenencia al grupo, y las puntuaciones discriminantes. A la raíz cuadrada de este coeficiente se le denomina correlación canónica y puede expresarse como:

$$\eta = \sqrt{\frac{\lambda}{1+\lambda}}$$

En cuanto a la **selección de variables**, en las aplicaciones de análisis discriminante se dispone frecuentemente de observaciones de un número relativamente elevado de variables potencialmente discriminantes. Aunque en todos los desarrollos anteriores se ha considerado que se conocen *a priori* cuáles son las variables clasificadoras, en la práctica se impone, cuando el número de variables es elevado, aplicar un sistema que permita seleccionar las variables con más capacidad discriminante entre un conjunto de variables más amplio. En el análisis discriminante, al igual que en el análisis de la regresión, los tres métodos más conocidos para selección de variables son los siguientes: selección hacia adelante (*forward*), selección hacia atrás (*backward*) y selección paso a paso (*stepwise*). Este último combina las características de los otros dos y además es el que se aplica con mayor frecuencia. Los tres procedimientos enunciados son procedimientos de carácter iterativo.

En cuanto al **cálculo de probabilidades de pertenencia a una población**, las funciones discriminantes clasifican a los diferentes individuos en uno u otro grupo, pero no dan más información acerca de los individuos investigados. En muchas ocasiones es conveniente tener información complementaria a las puntuaciones discriminantes. Con estas puntuaciones se puede clasificar a cada individuo, pero es interesante disponer además de información sobre la probabilidad de su pertenencia a cada grupo, ya que ello permitiría realizar análisis más matizados, e incluir otras informaciones tales como la información *a priori* o los costes que implica una clasificación errónea. Para realizar este tipo de cálculos se suelen asumir las hipótesis $x_g \rightarrow N(\mu_g, \Sigma)$, pero considerando que se conocen los parámetros poblacionales. Esta forma de proceder ocasiona ciertos problemas de los que nos ocuparemos posteriormente.

El cálculo de probabilidades se puede realizar en el contexto de la Teoría de la decisión, que permite tener en cuenta tanto la probabilidad de pertenencia a un grupo, como los costes de una clasificación errónea.

La clasificación de los individuos se puede realizar utilizando el teorema de Bayes, que permite el cálculo de las probabilidades *a posteriori* a partir de estas probabilidades *a priori* y de la información muestral contenida en las puntuaciones discriminantes. Considerando el caso general de G grupos, el teorema de Bayes establece que la probabilidad *a posteriori* de pertenencia a un grupo g con una puntuación discriminante D ($Prob(g/D)$) es la siguiente:

$$Prob(g/D) = \frac{\pi_g \times Prob(D/g)}{\sum_{i=1}^G \pi_i \times Prob(D/i)}$$

En el segundo miembro aparecen las probabilidades *a priori* π_g y las probabilidades condicionadas $Prob(D/g)$. La probabilidad condicionada $Prob(D/g)$ se obtiene calculando la probabilidad de la puntuación observada suponiendo la pertenencia a un grupo g . Dado que el denominador del segundo miembro del cociente anterior es una constante, se utiliza también, de forma equivalente, la siguiente expresión:

$$Prob(g/D) \propto \pi_g \times Prob(D/g)$$

donde el símbolo \propto significa proporcionalidad.

La clasificación de cada individuo se puede realizar mediante la comparación de las probabilidades *a posteriori*. Así, se asignará un individuo al grupo para el cual sea mayor su probabilidad *a posteriori*. Aunque a partir de ahora solamente se tratará el caso de 2 grupos, se va presentar el cálculo de probabilidades de forma que sea fácilmente generalizada al caso de G grupos.

El cálculo de probabilidades se va realizar bajo tres supuestos diferentes: cálculo de probabilidades sin información *a priori*, cálculo de probabilidades con información *a priori* y cálculo de probabilidades con información *a priori* y costes.

En cuanto al **cálculo de probabilidades sin información *a priori***, se considera que no existe conocimiento previo de las probabilidades de pertenencia a cada grupo. Cuando no existe dicha información, se adopta el supuesto de que la probabilidad de pertenencia a ambos grupos es la misma, es decir, se adopta el supuesto de que $\pi_I = \pi_{II}$. Esto implica que estas probabilidades *a priori* no afectan a los cálculos de las probabilidades *a posteriori*.

Bajo las hipótesis $x_g \rightarrow N(\mu_g, \Sigma)$, la probabilidad de pertenencia a cada grupo, dada la puntuación discriminante obtenida, viene dada como sigue:

$$Prob(g/D) = \frac{e^{F_g}}{e^{F_I} + e^{F_{II}}} \quad g = I, II$$

$$F_I = a_{I,1}X_1 + a_{I,2}X_2 + \cdots + a_{I,k}X_k - C_I$$

$$F_{II} = a_{II,1}X_1 + a_{II,2}X_2 + \cdots + a_{II,k}X_k - C_{II}$$

Un individuo se clasifica en el grupo para el que la probabilidad $Prob(g/D)$ sea mayor. Este criterio implica que un individuo se clasificará en el grupo I si $F_I > F_{II}$. Aplicando este criterio se llega a los mismos resultados que aplicando la función discriminante de Fisher. Esto implica que el punto de corte C que habíamos definido mediante:

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2}$$

sigue siendo aplicable con este nuevo enfoque.

Existe otro método para clasificar, que consiste en minimizar la probabilidad de clasificación errónea. Denominando $Prob(I/II)$ la probabilidad de clasificar a un individuo en la población I perteneciendo realmente a la II y $Prob(II/I)$ la probabilidad de clasificar a un individuo en la población II perteneciendo a la I, la probabilidad total de clasificación errónea es igual a:

$$Prob(II/I) + Prob(I/II)$$

Minimizando esta probabilidad, bajo las hipótesis $x_g \rightarrow N(\mu_g, \Sigma)$, se obtiene también como punto de corte el valor de C dado anteriormente.

En cuanto al **cálculo de probabilidades con información a priori**, en ocasiones se dispone de información de la probabilidad *a priori* sobre pertenencia de un individuo a cada uno de los grupos. Para tener en cuenta este tipo de información vamos a introducir probabilidades *a priori* en nuestro análisis.

Cuando se utilizan probabilidades *a priori* los individuos, o casos, se clasifican en el grupo para el que la probabilidad *a posteriori* sea mayor. De acuerdo con la hipótesis $x_g \rightarrow N(\mu_g, \Sigma)$, la probabilidad *a posteriori* de pertenencia a cada grupo se calcula como sigue:

$$Prob(g/D) = \frac{\pi_g e^{F_g}}{\pi_I e^{F_I} + \pi_{II} e^{F_{II}}} \quad g = I, II$$

Con este criterio se clasifica a un individuo en el grupo I si:

$$F_I \ln(\pi_I) > F_{II} \ln(\pi_{II})$$

El punto de corte discriminante C_g que habíamos definido mediante:

$$C_g = \frac{\bar{D}_I + \bar{D}_{II}}{2} - \ln \frac{\pi_{II}}{\pi_I}$$

La *radio* de probabilidades *a priori* debe establecerse de forma que el punto de corte se desplace hacia el grupo con menor probabilidad *a priori*. Al desplazar el punto de corte de esta forma, se tenderá a clasificar una proporción menor de individuos en el grupo con menor probabilidad *a priori*. Cuando las dos probabilidades *a priori* son igual a 0,5, entonces $C_g = C$.

Si se introducen probabilidades *a priori*, la probabilidad total de clasificación errónea es igual a:

$$\pi_I \times \text{Prob}(II/I) + \pi_{II} \times \text{Prob}(I/II)$$

Como puede verse, cada probabilidad de clasificación errónea va multiplicada por la probabilidad *a priori* del grupo real de pertenencia. Bajo las hipótesis estadísticas $x_g \rightarrow N(\mu_g, \Sigma)$, se obtiene que el punto de corte es C_g .

En cuanto al *cálculo de probabilidades con información a priori y consideración de costes*, la novedad está en que ahora se considera el coste que una clasificación errónea puede tener. En muchas aplicaciones el coste de clasificación errónea puede diferir para cada uno de los grupos. Cuando se introducen costes de clasificación no puede hablarse ya de cálculo de probabilidades *a posteriori*. No obstante se puede obtener un criterio para clasificar minimizando el coste total de clasificación errónea. Este coste total viene dado por la siguiente expresión:

$$\pi_I \times \text{Prob}(II/I) \times \text{Coste}(II/I) + \pi_{II} \times \text{Prob}(I/II) \times \text{Coste}(I/II)$$

Como puede verse en esta expresión, cada probabilidad va multiplicada por el coste en que se incurre. Cuando se minimiza esta expresión bajo la hipótesis $x_g \rightarrow N(\mu_g, \Sigma)$, el punto de corte discriminante C_{gc} es el siguiente:

$$C_{g,c} = \frac{\bar{D}_I + \bar{D}_{II}}{2} - \ln \frac{\pi_{II} \times \text{Coste}(I/II)}{\pi_I \times \text{Coste}(II/I)}$$

En todos los desarrollos anteriores se ha supuesto que las probabilidades son conocidas. En la práctica, sin embargo, se utilizan estadísticos muestrales en su lugar. El empleo de estadísticos muestrales tiene como consecuencia que se subestime la probabilidad de clasificación errónea, sometiéndose por lo tanto sesgos sistemáticos en la clasificación. Para disminuir estos sesgos se han propuesto, entre otros, dos procedimientos alternativos.

El primer procedimiento consiste en dividir la muestra total en dos submuestras, utilizando la primera muestra para estimar la función discriminante, mientras que la segunda se utiliza para su validación. Así, la potencia discriminante de la función vendrá determinada por el porcentaje de individuos clasificados correctamente en esta segunda submuestra.

El segundo procedimiento consiste en excluir un individuo del grupo I , calcular la función discriminante, y clasificar después al individuo que se ha excluido. Haciendo lo mismo con el resto de los individuos del grupo I , se estima la $Prob(II/I)$ con el porcentaje de individuos que han sido clasificados en el grupo II . Procediendo de la misma forma con los individuos del grupo II , se estima la $Prob(I/II)$. A este segundo procedimiento se le conoce con la denominación *jackknife*.

CLASIFICACIÓN CON MÁS DE DOS GRUPOS

En un caso general de análisis discriminante con G grupos ($G > 2$) llamado ***análisis discriminante múltiple***, el número máximo de ejes discriminantes que se pueden obtener viene dado por $\min(G-1, k)$. Por lo tanto pueden obtenerse hasta $G-1$ ejes discriminantes, si el número de variables explicativas k es mayor o igual que $G-1$, hecho que suele ser siempre cierto, ya que en las aplicaciones prácticas el número de variables explicativas suele ser grande.

Cada una de las funciones discriminantes D_i se obtiene como función lineal de las k variables explicativas X , es decir:

$$D_i = u_{i1}X_1 + u_{i2}X_2 + \cdots + u_{ik}X_k \quad i=1,2,\dots,G-1$$

Los $G-1$ ejes discriminantes vienen definidos respectivamente por los vectores $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{G-1}$ definidos mediante las siguientes expresiones:

$$\mathbf{u}_1 = \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1k} \end{bmatrix} \quad \mathbf{u}_2 = \begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2k} \end{bmatrix} \quad \dots \quad \mathbf{u}_{G-1} = \begin{bmatrix} u_{G-11} \\ u_{G-12} \\ \vdots \\ u_{G-1k} \end{bmatrix}$$

Para la obtención del primer eje discriminante, al igual que en caso de dos grupos, se maximiza λ_1 , donde:

$$\lambda_1 = \frac{\mathbf{u}_1' F \mathbf{u}_1}{\mathbf{u}_1' W \mathbf{u}_1}$$

La solución a este problema se obtiene derivando λ_1 respecto de u e igualando a cero, es decir:

$$\frac{\partial \lambda_1}{\partial u_1} = \frac{2F\mathbf{u}_1(\mathbf{u}_1' W \mathbf{u}_1) - 2W\mathbf{u}_1(\mathbf{u}_1' F \mathbf{u}_1)}{(\mathbf{u}_1' W \mathbf{u}_1)^2} = 0 \Rightarrow 2F\mathbf{u}_1(\mathbf{u}_1' W \mathbf{u}_1) - 2W\mathbf{u}_1(\mathbf{u}_1' F \mathbf{u}_1) = 0$$

De donde:

$$\frac{2F\mathbf{u}_1}{2W\mathbf{u}_1} = \frac{\mathbf{u}_1' F \mathbf{u}_1}{\mathbf{u}_1' W \mathbf{u}_1} = \lambda_1 \Rightarrow F\mathbf{u}_1 = W\mathbf{u}_1 \lambda_1 \Rightarrow W^{-1}F\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

Por lo tanto, la ecuación para la obtención del primer eje discriminante $W^{-1}F\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ se traduce en la obtención de un vector propio \mathbf{u}_1 asociado a la matriz no simétrica $W^{-1}F$.

De los valores propios λ_i que se obtienen al resolver la ecuación $W^{-1}F\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ se retiene el mayor, ya que precisamente λ_1 es la ratio que queremos maximizar y \mathbf{u}_1 es el vector propio asociado al mayor valor propio de la matriz $W^{-1}F$.

Dado que λ_1 es la ratio a maximizar, nos medirá, una vez calculado, el poder discriminante del primer eje discriminante. Como estamos en un caso general de análisis discriminante con G grupos ($G > 2$), el número máximo de ejes discriminantes que se pueden obtener viene dado por $\min(G-1, k)$. Por lo tanto pueden obtenerse hasta $G-1$ ejes discriminantes, si el número de variables explicativas k es mayor o igual que $G-1$, hecho que suele ser siempre cierto, ya que en las aplicaciones prácticas el número de variables explicativas suele ser grande.

El resto de los ejes discriminantes vendrán dados por los vectores propios asociados a los valores propios de la matriz $W^{-1}F$ ordenados de mayor a menor. Así, el segundo eje discriminante tendrá menos poder discriminatorio que el primero, pero más que cualquiera de los restantes.

Como la matriz $W^{-1}F$ no es simétrica, los ejes discriminantes no serán en general ortogonales (perpendiculares entre sí).

Podemos concluir que los ejes discriminantes son las componentes de los vectores propios normalizados asociados a los valores propios de la matriz $W^{-1}F$ ordenados en sentido decreciente (a mayor valor propio mejor eje discriminante).

En cuanto a los **contrastos de significación**, en el análisis discriminante múltiple se plantean contrastes específicos para determinar si cada uno de los valores λ_i que se obtienen al resolver la ecuación $W^{-1}Fu = \lambda u$ es estadísticamente significativo, es decir, para determinar si contribuye o no a la discriminación entre los diferentes grupos.

Este tipo de contrastes se realiza a partir del estadístico V de Barlett, que es una función de la Λ de Wilks y que se aproxima a una chi-cuadrado. Su expresión es la siguiente:

$$V = -\left\{ n - 1 - \frac{k + G}{2} \right\} \ln(\Lambda) \rightarrow \chi^2_{k(G-1)} \quad \Lambda = \frac{|W|}{|T|}$$

La hipótesis nula de este contraste es $H_0 : \mu_1 = \mu_2 = \dots = \mu_G$, y ha de ser rechazada para que se pueda continuar con el análisis discriminante, porque en caso contrario las variables clasificadoras utilizadas no tendrían poder discriminante alguno.

No olvidemos que W era la matriz suma de cuadrados y productos cruzados intragrupos en el análisis de la varianza múltiple y T era la matriz suma de cuadrados y productos cruzados total.

También existe un **estadístico de Barlett para contrastación secuencial**, que se elabora como sigue:

$$\frac{I}{A} = \frac{|T|}{|W|} = |W|^{-1}|T| = |W^{-1}T| = |W^{-1}(W + F)| = |I + W^{-1}F|$$

Pero como el determinante de una matriz es igual al producto de sus valores propios, se tiene que:

$$\frac{1}{\Lambda} = (1 + \lambda_1)(1 + \lambda_2) \cdots (1 + \lambda_{G-1})$$

Esta expresión puede sustituirse en la expresión del estadístico V vista anteriormente, para obtener la expresión alternativa siguiente para el estadístico de Barlett:

$$V = -\left\{ n - 1 - \frac{k + G}{2} \right\} \ln(\Lambda) = -\left\{ n - 1 - \frac{k + G}{2} \right\} \sum_{g=1}^{G-1} \ln(1 + \lambda_g) \rightarrow \chi^2_{k(G-1)}$$

Si se rechaza la hipótesis nula de igualdad de medias, al menos uno de los ejes discriminantes es estadísticamente significativo, y será el primero, porque es el que más poder discriminante tiene.

Una vez visto que el primer eje discriminante es significativo, se pasa a analizar la significatividad del segundo eje discriminante a partir del estadístico:

$$V = -\left\{ n - 1 - \frac{k + G}{2} \right\} \sum_{g=2}^{G-1} \ln(1 + \lambda_g) \rightarrow \chi^2_{(k-1)(G-2)}$$

De la misma forma se analiza la significatividad de sucesivos ejes discriminantes, pudiendo establecerse el estadístico V de Barlett genérico para contrastación secuencial de la significatividad del eje discriminante j-ésimo como:

$$V = -\left\{ n - 1 - \frac{k + G}{2} \right\} \sum_{g=j+1}^{G-1} \ln(1 + \lambda_g) \rightarrow \chi^2_{(k-j)(G-j-1)} \quad j = 0, 1, 2, \dots, G-2$$

En este proceso secuencial se van eliminando del estadístico V las raíces características que van resultando significativas, deteniendo el proceso cuando se acepte la hipótesis nula de no significatividad de los ejes discriminantes que queden por contrastar.

Como una medida descriptiva complementaria de este contraste se suele calcular el porcentaje acumulativo de la varianza después de la incorporación de cada nueva función discriminante.

ANÁLISIS DISCRIMINANTE CANÓNICO

Ya sabemos que en el análisis discriminante hay dos enfoques. El primero de ellos está basado en la obtención de funciones discriminantes de cálculo similar a las ecuaciones de regresión lineal múltiple y que es el que se ha tratado hasta ahora en este capítulo. El segundo emplea técnicas de correlación canónica y de componentes principales y se denomina *análisis discriminante canónico*.

Ya sabemos que el análisis en componentes principales es una técnica multivariante que persigue *reducir la dimensión de una tabla de datos excesivamente grande* por el elevado número de variables que contiene x_1, x_2, \dots, x_n y quedarse con unas cuantas variables C_1, C_2, \dots, C_p combinación de las iniciales (*componentes principales perfectamente calculables*) y que sinteticen la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables:

$$\begin{aligned}C_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\&\vdots \\C_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n\end{aligned}$$

Pero sólo se retienen las p componentes (componentes principales) que explican un porcentaje alto de la variabilidad de las variables iniciales (C_1, C_2, \dots, C_p). Se sabe que la primera componente C_1 tiene asociado el mayor valor propio de la matriz inicial de datos y que las sucesivas componentes C_2, \dots, C_p tienen asociados los siguientes valores propios en cuantía decreciente de su módulo. De esta forma, el análisis discriminante de dos grupos equivaldría al análisis en componentes principales con una sola componente C_1 . En este caso la única función discriminante canónica sería la ecuación de la componente principal $C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n$ y el valor propio asociado sería el poder discriminante. Para el análisis discriminante de tres grupos las funciones discriminantes canónicas serán las ecuaciones de las dos primeras componentes principales C_1 y C_2 , siendo su poder discriminante los dos primeros valores propios de la matriz de datos. De este modo, las componentes principales pueden considerarse como los sucesivos ejes de discriminación. Los coeficientes de la ecuación de cada componente principal, es decir, de cada eje discriminante, muestran el peso que cada variable aporta a la discriminación. No olvidemos que estos coeficientes están afectados por las escalas de medida, lo que indica que todas las variables deben presentar unidades parecidas, lo que se consigue estandarizando las variables iniciales antes de calcular las componentes principales.

SPSS Y LA CLASIFICACIÓN Y SEGMENTACIÓN MEDIANTE ANÁLISIS DISCRIMINANTE

PRINCIPIOS DEL ANÁLISIS DISCRIMINANTE

El análisis discriminante es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas. El objetivo esencial del análisis discriminante es utilizar los valores conocidos de las variables independientes para predecir con qué categoría de la variable dependiente se corresponden. Así por ejemplo, podremos predecir en qué categoría de riesgo crediticio se encuentra una persona, el éxito de un producto en el mercado, etc.

La expresión funcional del análisis discriminante es la siguiente:

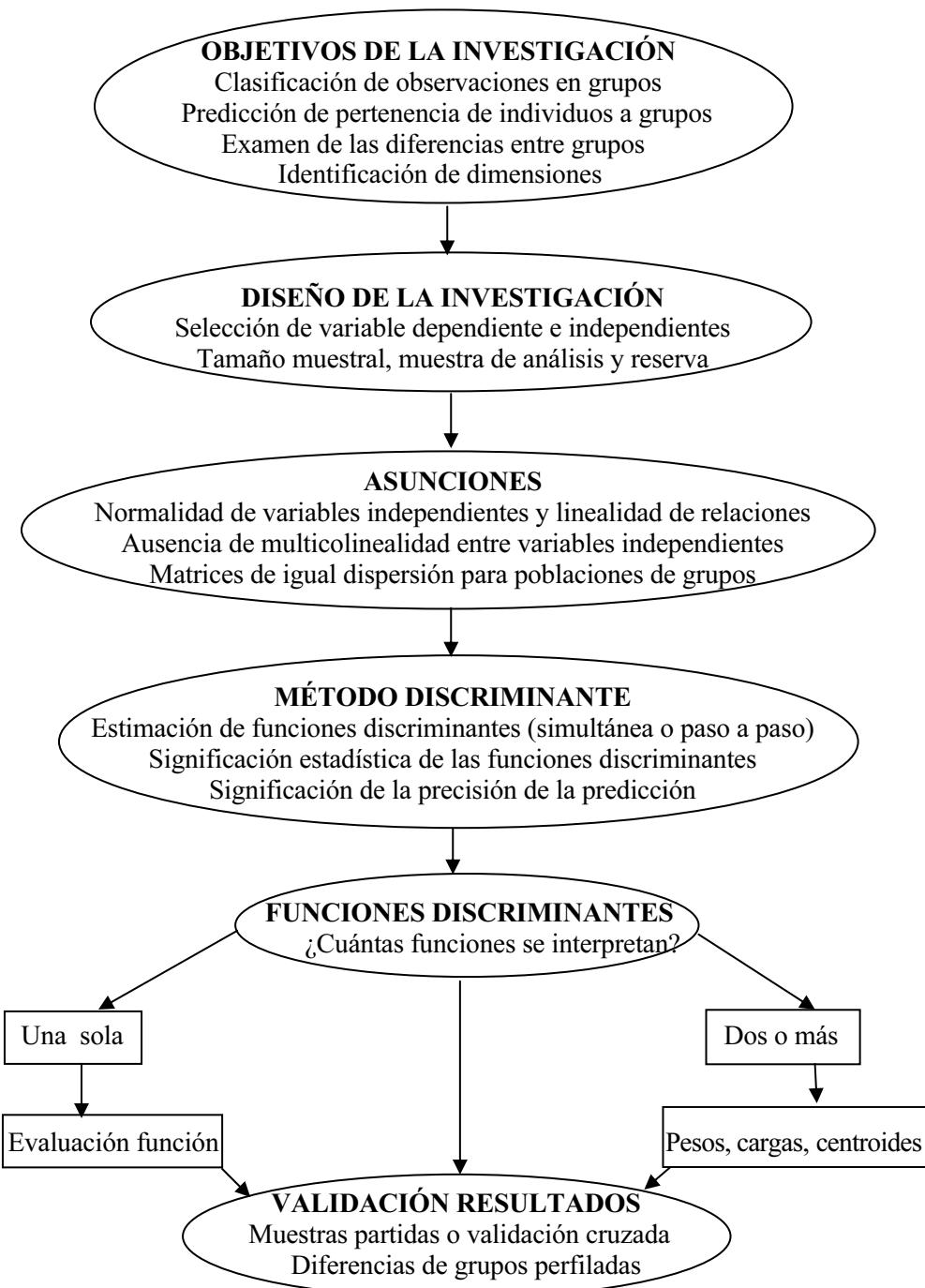
$$y = F(x_1, x_2, \dots, x_n)$$

La variable dependiente y es no métrica y las variables independientes son métricas.

Formalmente podríamos decir que el análisis discriminante es una técnica de clasificación que permite agrupar a los elementos de una muestra en dos o más categorías diferentes, predefinidas en una variable dependiente no métrica, en función de una serie de variables independientes métricas combinadas linealmente.

Pero al mismo tiempo, el análisis discriminante facilita el examen de las diferencias entre dos o más grupos, si atendemos a una serie de variables consideradas simultáneamente, permitiendo identificar dimensiones en función de las cuales difieren los grupos. Por lo tanto son propósitos del análisis discriminante la descripción de diferencias entre grupos y la predicción de pertenencia a los grupos en función de ciertas características conocidas para los sujetos dadas por las variables independientes métricas.

ESQUEMA GENERAL DEL ANÁLISIS DISCRIMINANTE



SPSS Y EL ANÁLISIS DISCRIMINANTE

A lo largo del Capítulo anterior hemos visto que el análisis discriminante resulta útil para las situaciones en las que se desea construir un modelo predictivo para pronosticar el grupo de pertenencia de un caso a partir de las características observadas de cada caso.

El procedimiento genera una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basada en combinaciones lineales de las variables predictoras que proporcionan la mejor discriminación posible entre los grupos. Las funciones se generan a partir de una muestra de casos para los que se conoce el grupo de pertenencia; posteriormente, las funciones pueden ser aplicadas a nuevos casos que dispongan de medidas para las variables predictoras pero de los que se desconozca el grupo de pertenencia. La variable de agrupación puede tener más de dos valores. Los códigos de la variable de agrupación han de ser números enteros y es necesario especificar sus valores máximo y mínimo. Los casos con valores fuera de estos límites se excluyen del análisis.

Para cada variable se obtendrán medias, desviaciones típicas y ANOVA univariado. Para cada análisis se obtendrán M de Box, matriz de correlaciones intra-grupos, matriz de covarianza intra-grupos, matriz de covarianza de los grupos separados y matriz de covarianza total. Para cada función discriminante canónica se obtendrán autovalores, porcentaje de varianza, correlación canónica, lambda de Wilks, G-cuadrado. Para cada paso se obtendrán probabilidades previas (*a priori*), coeficientes de la función de Fisher, coeficientes de función no tipificados y lambda de Wilks para cada función canónica.

Como ejemplo de aplicación del análisis discriminante podemos considerar el siguiente. Por término medio, las personas de los países de zonas templadas consumen más calorías por día que las de los trópicos, y una proporción mayor de la población de las zonas templadas vive en núcleos urbanos. Un investigador desea combinar esta información en una función para determinar cómo de bien un individuo es capaz de discriminar entre los dos grupos de países. El investigador considera además que el tamaño de la población y la información económica también pueden ser importantes. El análisis discriminante permite estimar los coeficientes de la función discriminante lineal, que tiene el aspecto de la parte derecha de una ecuación de regresión lineal múltiple. Es decir, utilizando los coeficientes *a*, *b*, *c* y *d*, la función es:

$$D = a * \text{clima} + b * \text{urbanos} + c * \text{población} + d * \text{producto interior bruto per capita}$$

Si estas variables resultaran útiles para discriminar entre las dos zonas climáticas, los valores de D serán diferentes para los países templados y para los tropicales. Si se utiliza un método de selección de variables por pasos, quizás no se necesite incluir las cuatro variables en la función.

Para realizar un análisis discriminante, elija en los menús *Analizar* → *Clasificar* → *Discriminante* (Figura 16-1) y seleccione las variables y las especificaciones para el análisis (Figura 16-2). Previamente es necesario cargar en memoria el fichero de nombre MUNDO mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo. Las variables independientes a considerar son: consumo diario de calorías (*calorías*), el logaritmo del PIB (*log_pib*), la población urbana (*urbana*) y el logaritmo de la población (*log_pob*). Como variable de agrupación usamos el clima (*clima*) con valores entre 5 y 8.

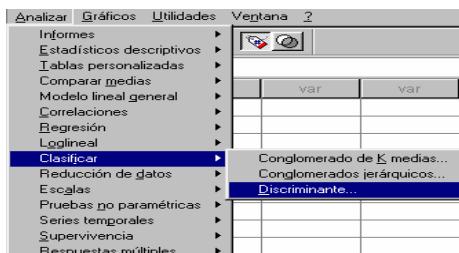


Figura 16-1

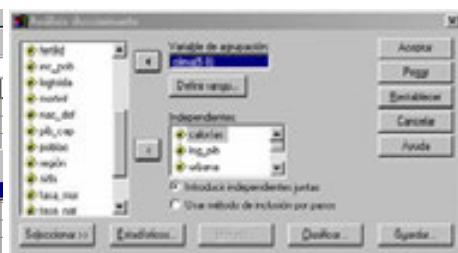


Figura 16-2

En cuanto a los datos, la variable de agrupación debe tener un número limitado de categorías distintas, codificadas como números enteros. Las variables independientes que sean nominales deben ser recodificadas a variables *dummy* o de contraste.

Los casos deben ser independientes. Las variables predictoras deben tener una distribución normal multivariada y las matrices de varianza-covarianza intra-grupos deben ser iguales en todos los grupos. Se asume que la pertenencia al grupo es mutuamente exclusiva (es decir, ningún caso pertenece a más de un grupo) y exhaustiva de modo colectivo (es decir, todos los casos son miembros de un grupo). El procedimiento es más efectivo cuando la pertenencia al grupo es una variable verdaderamente categórica; si la pertenencia al grupo se basa en los valores de una variable continua (por ejemplo, un cociente de inteligencia alto respecto a uno bajo), deberá considerar el uso de la regresión lineal para aprovechar la información más rica ofrecida por la propia variable continua.

El botón *Estadísticos* de la Figura 16-2 nos lleva a la pantalla de la Figura 16-3 en la que se establecen los estadísticos más relevantes relativos a las variables que ofrecerá el análisis. En el cuadro *Descriptivos* las opciones disponibles son: medias (que incluye las desviaciones típicas), ANOVAs univariados y M de Box (para igualdad de varianzas de grupos). En el cuadro *Coeficientes de la función* las opciones disponibles son: coeficientes de clasificación de Fisher y coeficientes no tipificados. En el cuadro *Matrices* se pueden elegir las matrices de coeficientes disponibles para las variables independientes: correlación intra-grupos, covarianza intra-grupos, covarianza de grupos separados y covarianza total.

El botón *Clasificar* de la Figura 16-2 nos lleva a la pantalla de la Figura 16-4 cuyo cuadro *Probabilidades previas* fija el uso de las probabilidades a priori para la clasificación. Puede especificarse que las probabilidades previas sean iguales para todos los grupos o dejar que los tamaños de grupo observados en la muestra determinen las probabilidades de la pertenencia al grupo. El cuadro *Mostrar* fija las opciones de presentación disponibles, que son: resultados por casos, tabla de resumen y clasificación dejando uno fuera. El botón *Reemplazar los valores perdidos con la media* permite sustituir la media de una variable independiente para un valor perdido, sólo durante la fase de clasificación. El cuadro *Usar matriz de covarianza* permite clasificar los casos utilizando una matriz de covarianza intra-grupos o una matriz de covarianza de los grupos separados. El cuadro *Gráficos* permite elegir entre las opciones de gráficos disponibles, que son: grupos combinados, grupos separados y mapa territorial.

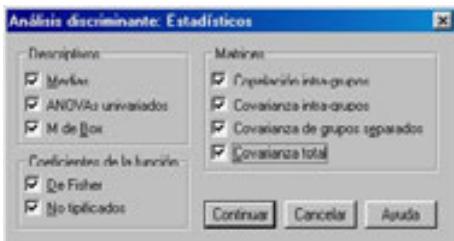


Figura 16-3

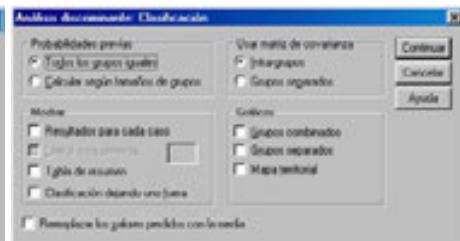


Figura 16-4

El botón *Definir rango* de la Figura 16-2 lleva a la pantalla de la Figura 16-5 y sirve para especificar los valores mínimo y máximo de la variable de agrupación para el análisis. Los casos con valores fuera de este rango no se utilizan en el análisis discriminante, pero sí se clasifican en uno de los grupos existentes a partir de los resultados que se obtengan en el análisis. Los valores mínimo y máximo deben ser números enteros.

El botón *Guardar* de la Figura 16-2 lleva a la pantalla de la Figura 16-6 y sirve para utilizar la posibilidad de añadir variables nuevas al archivo de datos activo. Las opciones disponibles son las de grupo de pertenencia pronosticado (una única variable), puntuaciones discriminantes (una variable para cada función discriminante en la solución) y probabilidades de pertenencia al grupo según las puntuaciones discriminantes (una variable para cada grupo). También se puede exportar información del modelo al archivo especificado. SmartScore y las próximas versiones de WhatIf? podrán utilizar este archivo.

El botón *Método* de la Figura 16-2 lleva a la pantalla de la Figura 16-7 y sirve para utilizar un método de inclusión por pasos. El cuadro *Método* permite seleccionar el estadístico que se va a utilizar para introducir o eliminar nuevas variables útiles para el análisis discriminante (no todas lo son). Las alternativas disponibles son: Lambda de Wilks (se eligen variables con el menor λ en la función discriminante), varianza no explicada (se buscan variables que la minimicen), distancia de Mahalanobis (se introducen variables que la maximicen), F de entrada y salida (se eligen las variables con estos valores altos) y V de Rao (a mayor cambio en V mejor discrimina la variable). Con la V de Rao se puede especificar el incremento mínimo de V para introducir una variable.

El cuadro *Criterios* permite utilizar las alternativas siguientes: usar valor de F y usar la probabilidad de F. Introduzca los valores para introducir y eliminar variables. El botón *Mostrar Resumen* de los pasos muestra los estadísticos para todas las variables después de cada paso. El botón F para distancias por parejas muestra una matriz de razones F por parejas para cada pareja de grupos. El botón *Aceptar* en la Figura 16-2 obtiene los resultados del análisis discriminante según se muestra en la Figura 16-8.



Figura 16-5

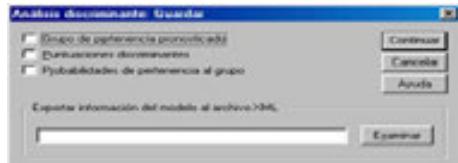


Figura 16-6



Figura 16-7

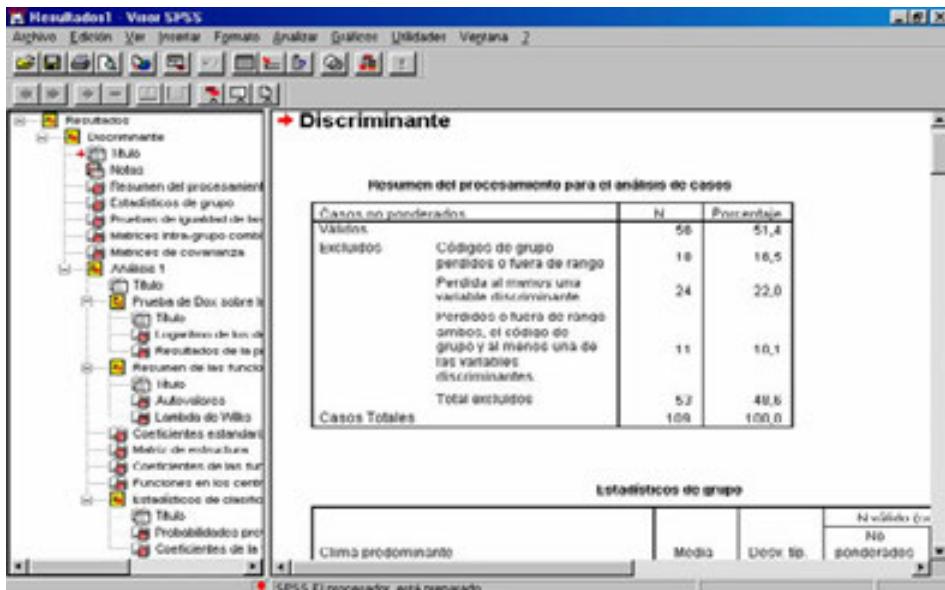


Figura 16-8

En la parte izquierda de la Figura 16-8 podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 16-9 a 16-17 se presentan los estadísticos por grupos de discriminación (figura 16-9), Prueba de igualdad de medias de los grupos (figura 16-10) y la matriz de correlaciones o covarianzas (figura 16-11), su inversa, el estadístico KMO, la prueba de Bartlett, la prueba de Box para la igualdad de matrices de covarianzas de las poblaciones de las que provienen los grupos (figura 16-12), los coeficientes discriminantes y de clasificación (figura 16-14 a 16-17), etc.

Estadísticos de grupo

Clima predominante		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
tropical	Ingesta diaria de calorías	2374,9286	308,8087	28	28,000
	Log(10) de PIB_CAP	2,9299	,4733	28	28,000
	Habitantes en ciudades (%)	42,9286	22,8602	28	28,000
	Log(10) de POBLAC	4,1704	,6117	28	28,000
mediterráneo	Ingesta diaria de calorías	2829,3333	542,2345	6	6,000
	Log(10) de PIB_CAP	3,3913	,8471	6	6,000
	Habitantes en ciudades (%)	53,8333	26,2558	6	6,000
	Log(10) de POBLAC	4,6255	,8552	6	6,000
templado	Ingesta diaria de calorías	3197,7727	533,8921	22	22,000
	Log(10) de PIB_CAP	3,7839	,6300	22	22,000
	Habitantes en ciudades (%)	65,8636	24,8927	22	22,000
	Log(10) de POBLAC	4,2729	,6584	22	22,000
Total	Ingesta diaria de calorías	2746,8750	578,6814	56	56,000
	Log(10) de PIB_CAP	3,3148	,7006	56	56,000
	Habitantes en ciudades (%)	53,1071	25,9557	56	56,000
	Log(10) de POBLAC	4,2594	,6591	56	56,000

Figura 16-9

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
Ingesta diaria de calorías	,545	22,158	2	53	,000
Log(10) de PIB_CAP	,666	13,310	2	53	,000
Habitantes en ciudades (%)	,825	5,621	2	53	,006
Log(10) de POBLAC	,957	1,194	2	53	,311

Matrices intra-grupo combinadas^a

	Ingesta diaria de calorías	Log(10) de PIB_CAP	Habitantes en ciudades (%)	Log(10) de POBLAC
Covarianza	Ingesta diaria de calorías	189259,303	213,939	6713,028
	Log(10) de PIB_CAP	213,939	,339	10,955
	Habitantes en ciudades (%)	6713,028	10,955	576,779
	Log(10) de POBLAC	-24,171	-7,961E-02	-3,057
Correlación	Ingesta diaria de calorías	1,000	,845	,643
	Log(10) de PIB_CAP	,845	1,000	,783
	Habitantes en ciudades (%)	,643	,783	1,000
	Log(10) de POBLAC	-,085	-,208	-,194

a. La matriz de covarianza tiene 53 grados de libertad

Figura 16-10

Matrices de covarianza ^a					
Clima predominante	Ingesta diaria de calorías	Log(10) de PIB_CAP	Habitantes en ciudades (%)	Log(10) de POBLAC	
tropical	Ingesta diaria de calorías	95382,810	118,644	4815,180	-15,657
	Log(10) de PIB_CAP	118,644	,224	8,165	-9,468E-02
	Habitantes en ciudades (%)	4815,180	8,165	522,587	-3,904
	Log(10) de POBLAC	-15,657	-9,468E-02	-3,904	,374
mediterráneo	Ingesta diaria de calorías	294018,267	414,027	12939,467	-185,059
	Log(10) de PIB_CAP	414,027	,718	20,856	-,143
	Habitantes en ciudades (%)	12939,467	20,856	689,367	-7,747
	Log(10) de POBLAC	-185,059	-,143	-7,747	,731
templado	Ingesta diaria de calorías	285040,755	288,820	7670,634	3,190
	Log(10) de PIB_CAP	288,820	,397	12,185	-4,504E-02
	Habitantes en ciudades (%)	7670,634	12,185	619,647	-,851
	Log(10) de POBLAC	3,190	-4,504E-02	-,851	,433
Total	Ingesta diaria de calorías	334872,148	364,349	10703,559	-,717
	Log(10) de PIB_CAP	364,349	,491	14,951	-5,389E-02
	Habitantes en ciudades (%)	10703,559	14,951	673,697	-2,387
	Log(10) de POBLAC	-7,717	-5,369E-02	-2,387	,434

a. La matriz de covarianza total presenta 55 grados de libertad.

Figura 16-11

En la figura 16-13 se observa que la primera función discriminante explica casi toda la variabilidad del modelo (95,9%), lo que concuerda con el hecho de que la lamdda de Wilks indica que sólo es significativa la primera función discriminante. La matriz de estructura de la figura 16-14 muestra que todas las variables (salvo la cuarta) tienen la mayor correlación con la primera función discriminante. También muestra los coeficientes estandarizados de las funciones discriminantes canónicas. Los coeficientes sin estandarizar se ven en la figura 16-15, así como los valores de las funciones en los centroides de los grupos (valores más distintos para la primera función ya que discrimina mejor). Las probabilidades previstas para los grupos se ven en la figura 16-16 (grupos equiprobables) y los coeficientes de la función de clasificación de Fisher se observan en la figura 16-17.

Prueba de Box sobre la igualdad de las matrices de covarianza		Resumen de las funciones canónicas discriminantes																			
Logaritmo de los determinantes		Autovalores																			
<table border="1"> <thead> <tr> <th>Clima predominante</th> <th>Rango</th> <th>Logaritmo del determinante</th> </tr> </thead> <tbody> <tr> <td>tropical</td> <td>4</td> <td>13,049</td> </tr> <tr> <td>mediterráneo</td> <td>4</td> <td>14,035</td> </tr> <tr> <td>templado</td> <td>4</td> <td>14,820</td> </tr> <tr> <td>Intra-grupos combinada</td> <td>4</td> <td>14,306</td> </tr> </tbody> </table>		Clima predominante	Rango	Logaritmo del determinante	tropical	4	13,049	mediterráneo	4	14,035	templado	4	14,820	Intra-grupos combinada	4	14,306	Función	Autovalor	% de varianza	% acumulado	Correlación canónica
Clima predominante	Rango	Logaritmo del determinante																			
tropical	4	13,049																			
mediterráneo	4	14,035																			
templado	4	14,820																			
Intra-grupos combinada	4	14,306																			
Los rangos y logaritmos naturales de los determinantes impresos son los de las matrices de covarianza de los grupos.		1	,885*	95,9	95,9	,685															
		2	,036*	4,1	100,0	,192															
a. Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.																					
Resultados de la prueba		Lambda de Wilks																			
M de Box	24,495	Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.															
F	Aprox .960	1 a la 2	,511	34,565	8	,000															
	g11 20	2	,963	1,931	3	,587															
	g12 768,109																				
	Sig. ,510																				
Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.																					

Figura 16-12

Figura 16-13

Coeficientes estandarizados de las funciones discriminantes canónicas			
	Función		
	1	2	
Ingesta diaria de calorías	1,107	-.459	
Log(10) de PIB_CAP	,005	,548	
Habitantes en ciudades (%)	-,196	-,145	
Log(10) de POBLAC	,160	1,014	

Matriz de estructura			
	Función		
	1	2	
Ingesta diaria de calorías	,972*	-,175	
Log(10) de PIB_CAP	,753*	-,164	
Habitantes en ciudades (%)	,488*	-,207	
Log(10) de POBLAC	,103	,966*	

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas
Variables ordenadas por el tamaño de la correlación con la función.

Figura 16-14

Coeficientes de las funciones canónicas discriminantes			
	Función		
	1	2	
Ingesta diaria de calorías	,003	-,001	
Log(10) de PIB_CAP	,008	,942	
Habitantes en ciudades (%)	-,008	-,006	
Log(10) de POBLAC	,243	1,543	
(Constante)	-7,619	-6,473	

Coeficientes no tipificados

Funciones en los centroides de los grupos			
	Función		
	1	2	
Clima predominante			
tropical	-,888	-4,56E-02	
mediterráneo	,293	,546	
templado	1,050	-9,07E-02	

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Figura 16-15

Estadísticos de clasificación			
Probabilidades previas para los grupos			
Clima predominante	Previstas	Casos utilizados en el análisis	
		No ponderados	Ponderados
tropical	,333	29	28,000
mediterráneo	,333	8	8,000
templado	,333	22	22,000
Total	1,000	59	59,000

Figura 16-16

Clima predominante	Clima predominante		
	tropical	mediterráneo	templado
Ingesta diaria de calorías	3,510E-03	5,901E-03	6,500E-03
Log(10) de PIB_CAP	15,681	16,247	15,654
Habitantes en ciudades (%)	-,204	,218	,220
Log(10) de POBLAC	11,211	12,510	11,712
(Constante)	-47,445	-60,071	-62,093

Figura 16-17

Ejercicio 16-1. Utilizando el fichero 16-1.sav realizar un análisis discriminante que clasifique los alumnos según los estudios que realizan, el hábitat en el que viven, el número de libros que leen al año, las horas semanales de televisión, la calificación media en los estudios y el número de hermanos que tienen.

Rellenamos la pantalla de entrada del procedimiento *Análisis discriminante* como se indica en la Figura 16-18. La variable dependiente será *Estudios* y las variables independientes del modelo serán *edad*, *hábitat*, *lect*, *califest* y *numher*. Las pantallas *Estadísticos* y *Opciones* del procedimiento se llenan como se indica en las Figuras 16-19 y 16-20. Al pulsar *Continuar* y *Aceptar* se obtiene la salida del procedimiento (Figuras 16-21 a 16-27).

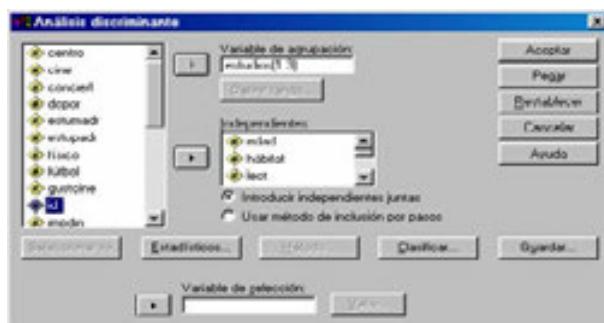


Figura 16-18

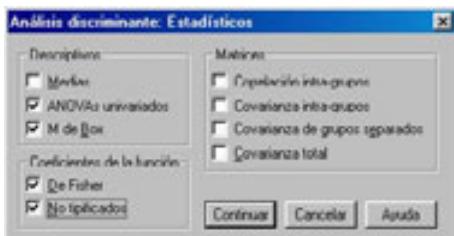


Figura 16-19

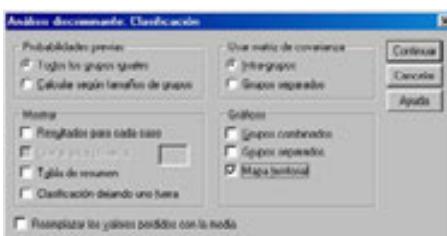


Figura 16-20

		Nº válido (según lista)	
		No ponderadas	Ponderadas
ESTUDIOS QUE CURSA			
EGB	EDAD	.99	.99
	HÁBITAT	.99	.99
	LIBROS LEÍDOS ANUALMENTE	.99	.99
	HORAS SEMANALES TV	.99	.99
	CALIFICACIÓN MEDIA EN ESTUDIOS	.99	.99
	Nº HERMANOS INCLUIDO SUJETO	.99	.99
BUP	EDAD	.80	.80
	HÁBITAT	.80	.80
	LIBROS LEÍDOS ANUALMENTE	.80	.80
	HORAS SEMANALES TV	.80	.80
	CALIFICACIÓN MEDIA EN ESTUDIOS	.80	.80
	Nº HERMANOS INCLUIDO SUJETO	.80	.80
FP	EDAD	.53	.53
	HÁBITAT	.53	.53
	LIBROS LEÍDOS ANUALMENTE	.53	.53
	HORAS SEMANALES TV	.53	.53
	CALIFICACIÓN MEDIA EN ESTUDIOS	.53	.53
	Nº HERMANOS INCLUIDO SUJETO	.53	.53
Total	EDAD	175	175.000
	HÁBITAT	175	175.000
	LIBROS LEÍDOS ANUALMENTE	175	175.000
	HORAS SEMANALES TV	175	175.000
	CALIFICACIÓN MEDIA EN ESTUDIOS	175	175.000
	Nº HERMANOS INCLUIDO SUJETO	175	175.000

Figura 16-21

Pruebas de igualdad de las medias de los grupos

	Lambdas de Wilks	F	gl1	gl2	Sig.
EDAD	.482	172,537	2	172	.000
HÁBITAT	.890	10,644	2	172	.000
LIBROS LEÍDOS ANUALMENTE	.808	6,881	2	172	.002
HORAS SEMANALES TV	.946	4,905	2	172	.008
CALIFICACIÓN MEDIA EN ESTUDIOS	.984	1,392	2	172	.251
Nº HERMANOS INCLUIDO SUJETO	.875	2,220	2	172	.112

Análisis 1

Prueba de Box sobre la igualdad de las matrices de covarianza

Logaritmo de los determinantes

ESTUDIOS QUE CURSA	Rango	Logaritmo del determinante
EGB	6	2,710
BUP	6	7,081
FP	6	6,718
Interc. grupos combinados	6	6,879

Los ratios y los logaritmos naturales de los determinantes contrastan con los de las matrices de covarianza de los resúmenes.

Resultados de la prueba

M de Box	Aprox.	168,737
F	gl1	3,748
	gl2	47
	Sig.	49043,875
	Sig.	.000

Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.

Resumen de las funciones canónicas discriminantes

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1,646*	94,6	94,6	.789
2	.0944	5,4	100,0	.293

* Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.

Lambdas de Wilks:

Contraste de las funciones	Lambdas de Wilks	Chi cuadrado	gl	Sig.
1 a la 2	.345	160,187	12	.000
2	.914	15,231	5	.009

Figura 16-22

Figura 16-23

Según la Figura 16-22, tanto las calificaciones medias como el número de hermanos son similares al 95% en los tres grupos de estudios (p-valor mayor que 0,05). Sin embargo, en las otras tres variables hay diferencias significativas en las medias de los distintos grupos. En la Figura 16-23 se contrasta la homoscedasticidad del modelo mediante el estadístico M de Box, cuyo p-valor cero impide aceptar la hipótesis nula de igualdad de covarianzas de los grupos de discriminación. Los p-valores de cuadro Lambda de Wilks certifican la significatividad de los dos ejes discriminantes, con lo que su capacidad explicativa será buena (separan bien grupos).

El cuadro *Autovalores* de la Figura 16-23 presenta los autovalores de las funciones canónicas discriminantes, que miden las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones dentro de los grupos. El autovalor de una función se interpreta como la parte de variabilidad total de la nube de puntos proyectada sobre el conjunto de todas las funciones atribuible a la función. Si su valor es grande, la función discriminará mucho. En cuanto a las correlaciones canónicas, miden las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones totales sin distinguir grupos. Si su valor es grande (próximo a 1) la dispersión será debida a las diferencias entre grupos, y por tanto la función discriminará mucho.

En El cuadro *Autovalores* se observa también que los valores de la correlación canónica decrecen $0,789 > 0,293$ y la primera función discrimina más que la segunda. La primera función discrimina bien porque su correlación canónica es casi 0,8, pero la segunda debiera ser un poco mejor. Con los autovalores ocurre lo mismo $1,646 > 0,094$. La primera función explicaría un total del 94,6% de la variabilidad total, mientras que la segunda explica el restante 5,4%. La primera función es la que va a dar prácticamente la clasificación, mientras que la segunda aporta poca información, aunque ya lo hemos visto con la Lambda de Wilks que es significativa.

Coeficientes de las funciones canónicas discriminantes		
	Función	
	1	2
EDAD	,849	,181
HÁBITAT	-,570	1,303
LIBROS LEÍDOS		
ANUALMENTE	-,027	,141
HORAS SEMANALES TV	,004	,013
CALIFICACIÓN MEDIA EN ESTUDIOS	,091	,021
Nº HERMANOS		
INCLUIDO SUJETO	-,177	,315
(Constante)	-12,278	-8,290

Coeficientes no tipificados

Funciones en los centroides de los grupos		
	Función	
	1	2
ESTUDIOS QUE CURSA		
EGB	-2,375	9,453E-03
BUP	,662	-,278
FP	,711	,429

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Figura 16-24

Coeficientes estandarizados de las funciones discriminantes canónicas		
	Función	
	1	2
EDAD	,995	,212
HÁBITAT	-,263	,601
LIBROS LEÍDOS		
ANUALMENTE	-,161	,842
HORAS SEMANALES TV	,051	,186
CALIFICACIÓN MEDIA EN ESTUDIOS	,110	,026
Nº HERMANOS		
INCLUIDO SUJETO	-,251	,445

Matriz de estructura		
	Función	
	1	2
EDAD	,927*	,342
CALIFICACIÓN MEDIA EN ESTUDIOS	,096*	,096
HÁBITAT	-,225	,658*
LIBROS LEÍDOS		
ANUALMENTE	-,169	,569*
HORAS SEMANALES TV	,164	-,372*
Nº HERMANOS		
INCLUIDO SUJETO	,108	,268*

Figura 16-25

En la Figura 16-24 se ven los coeficientes de las funciones canónicas discriminantes, que indican que las funciones discriminantes se escriben como:

$$\begin{aligned} D1 &= -12,278 + 0,849 * EDAD - 0,57 * HABITAT - 0,27 * LECT + 0,004 * TV + 0,91 * CALIFEST - 1,77 * NUMHER \\ D2 &= -8,29 + 0,181 * EDAD + 1,303 * HABITAT + 0,141 * LECT + 0,013 * TV + 0,021 * CALIFEST + 0,315 * NUMHER \end{aligned}$$

El cuadro *Funciones* en los centroides de los grupos de la Figura 16-24 nos da una idea de cómo las funciones discriminan grupos. Si las medias de los dos grupos en cada función son muy parecidas la función no discrimina grupos. Se observa que la discriminación es buena tal y como ya había asegurado la Lambda de Wilks. No obstante, en ambas funciones la diferencia entre los grupos BUP y FP es más pequeña, sobre todo en la segunda función, lo que viene a refrendar su menor poder discriminatorio, tal y como ya habíamos visto anteriormente.

En la Figura 16-25 se ven los coeficientes de las funciones canónicas discriminantes con sus variables tipificadas, cuyas ecuaciones son:

$$\begin{aligned} D1 &= 0,995 * EDAD - 2,63 * HABITAT - 1,61 * LECT + 0,051 * TV + 0,11 * CALIFEST - 0,251 * NUMHER \\ D2 &= 0,212 * EDAD + 0,601 * HABITAT + 0,842 * LECT + 0,186 * TV + 0,26 * CALIFEST + 0,445 * NUMHER \end{aligned}$$

Con las funciones estandarizadas podemos apreciar qué variables influyen más en cada función discriminante. Por ejemplo, EDAD contribuye más a la discriminación en la primera función ($0,995 > 0,212$) y CLIFEST en la segunda ($0,11 < 0,26$). En la *Matriz de estructura* de la Figura 16-25 está marcada con asteriscos la mejor contribución de cada variable a cada función discriminante.

Estadísticos de clasificación		
Probabilidades previas para los grupos		
ESTUDIOS QUE CURSA	Previas	Casos utilizados en el análisis
		No ponderados Ponderados
EGB	,333	39 39,000
BUP	,333	83 83,000
FP	,333	53 53,000
Total	1,000	175 175,000

Figura 16-26

Coeficientes de la función de clasificación			
	ESTUDIOS QUE CURSA		
	EOB	BUP	FP
EDAD	9,880	12,406	12,576
HÁBITAT	8,973	6,867	7,760
LIBROS LEÍDOS ANUALMENTE	1,752	1,630	1,728
HORAS SEMANALES TV	,746	,753	,763
CALIFICACIÓN MEDIA EN ESTUDIOS	4,730	5,001	5,020
Nº HERMANOS INCLUIDO SUJETO	-2,579	-3,207	-2,994
(Constante)	-108,563	-140,903	-147,456

Figura 16-27

La Figura 16-27 presenta los coeficientes de las tres funciones lineales discriminantes de Fisher, que *se pueden utilizar directamente para clasificar a los individuos futuros, previo cálculo su puntuación en cada uno de los grupos usando las funciones discriminantes con estos coeficientes de Fisher. Cada individuo se clasificará en el grupo en el que haya alcanzado una puntuación más elevada*. La Figura 16-26 presenta las probabilidades a priori, que se utilizan también para clasificar a los estudiantes en grupos. Como hemos utilizado la opción por defecto de que todas las probabilidades iniciales de pertenencia sean iguales (botón *Todos los grupos iguales* del campo *Probabilidades previas* de la Figura 16-20), cada individuo tiene *a priori* la misma probabilidad 1/3 de pertenecer a cada uno de los grupos.

Ejercicio 16-2. Utilizando el fichero 16-2.sav y el método de inclusión por pasos, realizar un análisis discriminante que clasifique los individuos en grupos dependiendo del tipo de cine que les guste (amor, humor, violencia o sexo) registrado en la variable TIPOCINE, según la calificación media en los estudios (CALIFEST), el número de veces que anualmente van al cine (CINE), su edad (EDAD), el número de libros que leen al año (LECT), la paga semanal (PAGA), las horas semanales de televisión (TV) y el nivel de rechazo a la violencia que tienen (VIOLEN).

Rellenamos la pantalla de entrada del procedimiento *Análisis discriminante* como se indica en la Figura 16-28. La variable dependiente será *Tipocine* y las variables independientes del modelo serán *califest*, *cine*, *edad*, *lect*, *paga*, *tv* y *violenc*. Las pantallas *Estadísticos*, *Clasificar*, *Guadrar* y *Método* se llenan como se indica en las Figuras 16-29 a 16-32. Al pulsar *Continuar* y *Aceptar* se obtiene la salida del procedimiento. La figura 16-33 indica que hay 165 casos válidos en el análisis y que se han excluido 10 por las diversas causas que se exponen. La figura 16-34 muestra las pruebas de igualdad de medias de las variables independientes en los 4 grupos discriminantes (valores de la variable dependiente). Se ve que se acepta la igualdad de medias de las variables *paga*, *califest*, *lect* y *tv* en los 4 grupos (*p*-valores mayores que 0,05) y se rechaza la igualdad de medias para las otras tres *cine*, *violenc* y *edad*, que son las posibles para discriminar.



Figura 16-28

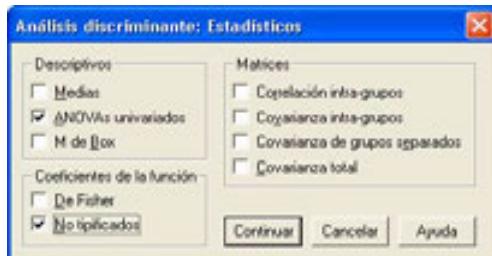


Figura 16-29

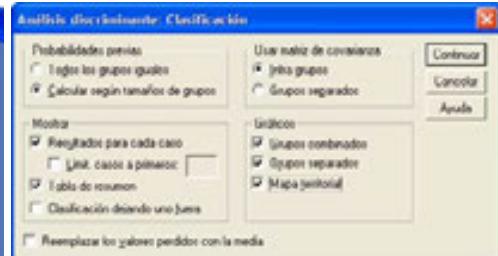


Figura 16-30

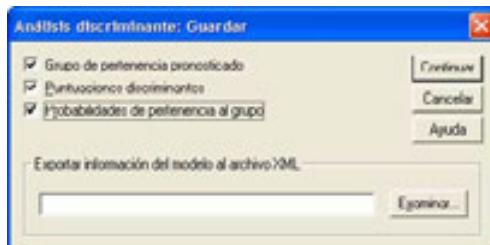


Figura 16-31

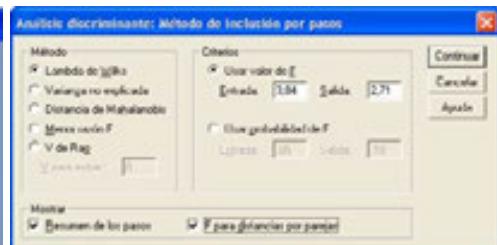


Figura 16-32

Resumen del procesamiento para el análisis de casos		
	N	Porcentaje
Casos no ponderados		
Válidos	165	94,3
Excluidos		
Códigos de grupo perdidos o fuera de rango	1	,6
Perdida al menos una variable discriminante	8	4,6
Perdidos o fuera de rango ambos, el código de grupo y al menos una de las variables discriminantes.	1	,6
Total excluidos	10	5,7
Casos Totales	175	100,0

Figura 16-33

Pruebas de igualdad de las medias de los grupos					
	Lambda de Wilks	F	gl1	gl2	Sig.
PAGA SEMANAL EN PTAS/M100	,992	,431	3	161	,731
EDAD	,885	6,980	3	161	,000
CALIFICACIÓN MEDIA EN ESTUDIOS	,984	,882	3	161	,452
LIBROS LEÍDOS ANUALMENTE	,981	1,034	3	161	,379
ASISTENCIA ANUAL AL CINE	,944	3,195	3	161	,025
HORAS SEMANALES TV	,998	,134	3	161	,940
NIVEL DE RECHAZO A LA VIOLENCIA	,433	70,149	3	161	,000

Figura 16-34

En el proceso de análisis discriminante se buscan funciones discriminantes a partir de las variables independientes para clasificar a los individuos según los valores de la variable dependiente. Por ello, inicialmente se seleccionan las variables independientes que más discriminan (que proporcionen los centros de los grupos muy distintos entre sí y muy homogéneos dentro de sí). Las figuras 16-35 y 16-37 nos muestran que las variables introducidas para discriminar en el modelo son definitivamente *violen* y *edad*. En la etapa 1 se seleccionó *violen* y en la etapa 2 se seleccionó *edad*. Los valores de la lambda de Wilks de la figura 16-37 (0,433 y 0,386) no son muy pequeños (no son próximos a cero) por lo que es posible que los grupos no estén claramente separados. Los p-valores de cuadro Lambda de Wilks y los estadístico F exacta (figura 16-38) certifican la significatividad de dos ejes discriminantes, con lo que su capacidad explicativa será buena (separan bien grupos). Luego el modelo formado por las dos variables es significativo (p-valores nulos). Para describir las dos funciones discriminantes canónicas puede usarse los coeficientes estandarizados $D1=-0,011edad+1,001violen$ y $D2=-1,004edad-0,82violen$ (figura 16-36) o sin estandarizar $D1=4,272-0,006edad+3,535violen$ y $D2=-8,832+0,583edad-0,290violen$ (figura 16-41). Se ve que *violen* contribuye más a la primera función ($1,001 > 0,82$) y *edad* a la segunda ($1,004 > 0,011$). En la matriz de estructura (figura 16-40) se fija este resultado.

Variables en el análisis

Paso	Tolerancia	F para eliminar	Lambda de Wilks
1 NIVEL DE RECHAZO A LA VIOLENCIA	1,000	70,149	
2 NIVEL DE RECHAZO A LA VIOLENCIA EDAD	,991	68,861	,885

Figura 16-35

Coeficientes estandarizados de las funciones discriminantes canónicas		
	Función	
	1	2
EDAD	-,011	1,004
NIVEL DE RECHAZO A LA VIOLENCIA	1,001	-,082

Figura 16-36

Variables introducidas/eliminadas ^{a,b,c,d}										
Paso	Introducidas	Lambda de Wilks						F exacta		
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.	
1	NIVEL DE RECHAZO A LA VIOLENCIA	,433	1	3	161,000	70,149	3	161,000	,000	
	EDAD	,386	2	3	161,000	32,484	6	320,000	,000	

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

a. El número máximo de pasos es 14.
b. La F parcial mínima para entrar es 3,84.
c. La F parcial máxima para eliminar es 2,71.
d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

Figura 16-37

Lambda de Wilks

Paso	Número de variables	Lambda	F exacta						
			gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	1	,433	1	3	161	70,149	3	161,000	,000
2	2	,386	2	3	161	32,484	6	320,000	,000

Figura 16-38

En la figura 16-39 se observa que la primera función discriminante explica casi toda la variabilidad del modelo (91,5%) mientras que la segunda sólo explica el 8,5%, aunque según los p-valores de la lambda de Wilks son significativas las dos funciones discriminantes. La matriz de estructura de la figura 16-40 muestra que las tres primeras variables tienen la mayor correlación con la primera función discriminante (sólo se emplea en el análisis *violenc*) y las tres últimas están más correladas con la segunda función discriminante (sólo se emplea en el análisis *edad*). En la figura 16-39 se observa que los valores de la correlación canónica decrecen $0,753 > 0,330$, con lo que la primera función discrimina más que la segunda. Con los auvalores ocurre lo mismo $1,307 > 1,22$. La primera función es la que va a dar prácticamente la clasificación, mientras que la segunda aporta poca información, aunque ya lo hemos visto con la Lambda de Wilks que es significativa. El cuadro *Funciones en los centroides de los grupos* de la Figura 16-41 nos da una idea de cómo las funciones discriminan grupos. Si las medias de los cuatro grupos en cada función son muy parecidas la función no discrimina grupos. Se observa que la discriminación es buena para las dos funciones tal y como ya había asegurado la Lambda de Wilks.

Resumen de las funciones canónicas discriminantes					
Autovalores					
Función	Autovalor	% de varianza	% acumulado	Correlación canónica	
1	1,307 ^a	91,5	91,5	,753	
2	,122 ^a	8,5	100,0	,330	

a. Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.

Lambda de Wilks					
Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.	
1 a la 2	,386	153,163	6	,000	
2	,891	18,556	2	,000	

Figura 16-39

Los individuos se clasifican en los cuatro grupos de acuerdo a las probabilidades que tienen a priori de pertenecer a los mismos (figura 16-42). Pero una vez conocidas las puntuaciones discriminantes (valores de las funciones discriminantes para cada individuo), cada individuo se clasificará en el grupo en que tenga mayor probabilidad a posteriori de pertenecer según sus puntuaciones discriminantes. La tabla *Resultados de la clasificación o matriz de confusión* de la figura 16-43 muestra los casos en total que están correcta o incorrectamente clasificados (75,1% correctos). Se muestran también tantos por ciento en cada grupo y en el total junto con el número de casos que se han clasificado en cada nivel.

En la tabla de *estadísticos por casos* de la figura 16-44 se observan el grupo real y el pronosticado (para *grupo mayor* y *segundo grupo mayor*) al que pertenece cada individuo (sólo los 30 primeros). ***Un individuo se clasifica en el grupo en el que su pertenencia tiene una mayor probabilidad a posteriori.*** Cuando el grupo real en que cae el individuo y el pronosticado en *grupo mayor* no coinciden, hay un error de clasificación del individuo. En la columna de *segundo grupo mayor* se observan los grupos a que pertenece cada individuo en segundo lugar en sentido probabilística (pero el importante es el *grupo mayor*). Las dos últimas columnas de la tabla de *estadísticos por casos* de la figura 16-44 muestran las puntuaciones discriminantes de los individuos para las dos funciones discriminantes. Los casos que tengan puntuaciones discriminantes similares se situarán próximos en los grupos de discriminación. No obstante, son más útiles las puntuaciones en los centroides de los grupos (figura 16-41) ya que determinan su posición en el espacio discriminante. La puntuación de un centroide se determina sustituyendo las variables de la ecuación discriminante por los valores medios de estas variables en el grupo. ***Una observación futura se clasificará en el grupo cuyo centroide esté más cerca de la puntuación discriminante de la observación según la función discriminante considerada. Lo ideal sería clasificar la observación en el mismo grupo según las dos funciones discriminantes.***

Matriz de estructura		
	Función	
	1	2
NIVEL DE RECHAZO A LA VIOLENCIA	1,000*	,011
CALIFICACIÓN MEDIA EN ESTUDIOS ^a	,099*	,073
ASISTENCIA ANUAL AL CINE ^a	-,018*	,013
EDAD	,082	,997*
HORAS SEMANALES TV ^a	,048	,144*
LIBROS LEÍDOS ANUALMENTE ^a	-,082	-,106*
PAGA SEMANAL EN PTAS/100 ^a	,000	,095*

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas
Variables ordenadas por el tamaño de la correlación con la función.

*. Mayor correlación absoluta entre cada variable y cualquier función discriminante.

a. Esta variable no se emplea en el análisis.

Figura 16-40

Coeficientes de las funciones canónicas discriminantes		
	Función	
	1	2
EDAD	-,006	,583
NIVEL DE RECHAZO A LA VIOLENCIA (Constante)	3,535	-,290
	-4,272	-8,832

Coeficientes no tipificados

Funciones en los centroides de los grupos		
TIPO DE PELÍCULA QUE TE GUSTA	Función	
	1	2
AMOR	-,791	,136
HUMOR	-,517	-,563
VIOLENCIA	1,833	,019
SEXO	-,404	,964

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Figura 16-41

Resumen del proceso de clasificación	
Procesados	175
Excluidos	0
Código de grupo perdido o fuera de rango	0
Perdida al menos una variable discriminante	1
Usados en los resultados	174

Probabilidades previas para los grupos			
TIPO DE PELÍCULA QUE TE GUSTA	Previás	Casos utilizados en el análisis	
		No ponderados	Ponderados
AMOR	,473	78	78,000
HUMOR	,206	34	34,000
VIOLENCIA	,273	45	45,000
SEXO	,048	8	8,000
Total	1,000	165	165,000

Figura 16-42

Resultados de la clasificación ^a							
Original	Recuento	TIPO DE PELÍCULA QUE TE GUSTA	Grupo de pertenencia pronosticado				Total
			AMOR	HUMOR	VIOLENCIA	SEXO	
		AMOR	81	1	1	0	83
		HUMOR	19	14	3	0	36
		VIOLENCIA	7	4	35	0	46
		SEXO	7	0	1	0	8
		Casos desagrupados	1	0	0	0	1
		%	AMOR	97,6	1,2	,0	100,0
			HUMOR	52,8	38,8	,0	100,0
			VIOLENCIA	15,2	8,7	,0	100,0
			SEXO	87,5	,0	12,5	100,0
			Casos desagrupados	100,0	,0	,0	100,0

a. Clasificados correctamente el 75,1% de los casos agrupados originales.

Figura 16-43

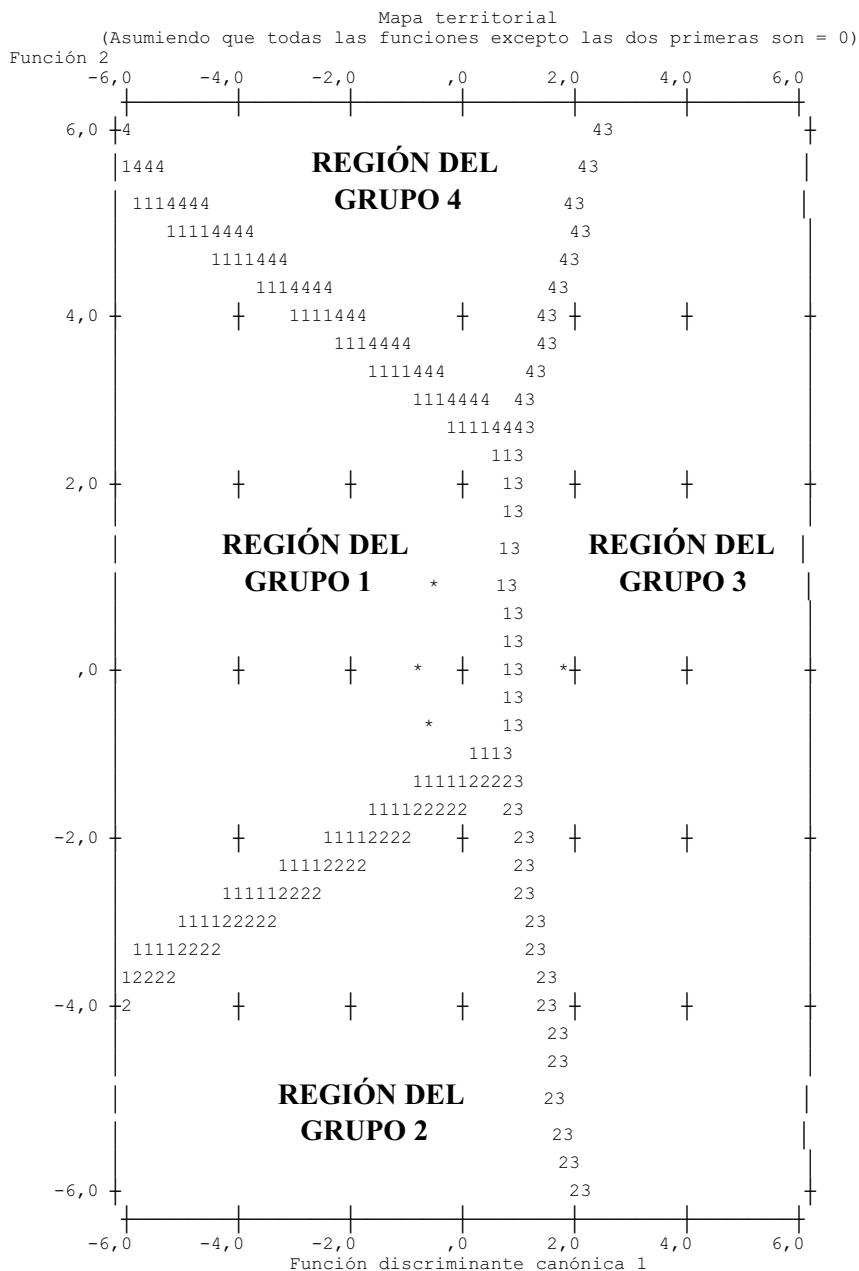
Estadísticos por casos

Número de casos	Grupo real	Grupo mayor			Segundo grupo mayor			Puntuaciones discriminantes			
		Grupo pronosticado	P(D>d G=g)	P(G=g D=d)	Distancia de Mahalanobis al cuadrado hasta el centroide	Grupo	P(G=g D=d)	Distancia de Mahalanobis al cuadrado hasta el centroide	F 1	F 2	
		p	g								
2	4	1(**)	,997	2	,716	,006	2	,222	,683	-,837	,199
3	1	1	,873	2	,654	,271	2	,306	,131	-,831	-,383
4	3	3	,619	2	,988	,960	1	,006	12,256	2,692	,492
5	1	1	,545	2	,572	1,215	2	,402	,257	-,824	-,966
6	2	1(**)	,469	2	,750	1,513	4	,137	,359	-,849	1,364
7	3	3	,684	2	,988	,759	2	,006	10,560	2,698	-,091
8	3	3	,684	2	,988	,759	2	,006	10,560	2,698	-,091
9	Desagr.	1	,873	2	,654	,271	2	,306	,131	-,831	-,383
10	1	1	,997	2	,716	,006	2	,222	,683	-,837	,199
11	2	3(**)	,539	2	,987	1,237	2	,008	10,389	2,704	-,673
12	2	1(**)	,469	2	,750	1,513	4	,137	,359	-,849	1,364
13	1	1	,997	2	,716	,006	2	,222	,683	-,837	,199
14	1	1	,873	2	,654	,271	2	,306	,131	-,831	-,383
15	1	1	,997	2	,716	,006	2	,222	,683	-,837	,199
16	2	1(**)	,997	2	,716	,006	2	,222	,683	-,837	,199
17	1	1	,997	2	,716	,006	2	,222	,683	-,837	,199
18	2	1(**)	,545	2	,572	1,215	2	,402	,257	-,824	-,966
19	1	1	,811	2	,749	,420	2	,155	1,915	-,843	,782
20	4	3(**)	,399	2	,987	1,840	1	,007	12,968	2,686	1,075
21	3	3	,539	2	,987	1,237	2	,008	10,389	2,704	-,673
22	2	3(**)	,619	2	,988	,960	1	,006	12,256	2,692	,492
23	3	3	,399	2	,987	1,840	1	,007	12,968	2,686	1,075
24	3	1(**)	,997	2	,716	,006	2	,222	,683	-,837	,199
25	1	1	,873	2	,654	,271	2	,306	,131	-,831	-,383
26	2	1(**)	,811	2	,749	,420	2	,155	1,915	-,843	,782
27	2	2	,280	2	,607	2,545	1	,381	5,138	-,812	-2,131
28	1	1	,997	2	,716	,006	2	,222	,683	-,837	,199
29	1	1	,811	2	,749	,420	2	,155	1,915	-,843	,782

** Caso mal clasificado

Figura 16-44

El **mapa territorial** que se muestra a continuación representa los valores de las puntuaciones en las funciones discriminantes canónicas (en abscisas se sitúan las puntuaciones en la función 1 y en ordenadas las puntuaciones en la función 2). La región del grupo 1 está delimitada por números 1 en el mapa, la del grupo 2 por el número 2, etc.



Símbolos usados en el mapa territorial

Simbol	Grupo	Etiqu
-----	-----	-----
1	1	AMOR
2	2	HUMOR
3	3	VIOLENCIA
4	4	SEXO
*		Indica un centroide de grupo

Cuando los casos o individuos están bien clasificados, su representación sobre el mapa territorial los sitúa en el territorio correspondiente al grupo. Cuando la discriminación es débil puede haber sujetos que caen fuera de su territorio y que estarían mal clasificados. Las líneas de números que separan una zona de otra delimitan las combinaciones de puntuaciones discriminantes en ambas funciones que conducen a la clasificación en cada grupo. *El mapa territorial también se utiliza para clasificar individuos futuros. Para ello se observan las puntuaciones del individuo en las funciones discriminantes consideradas y se observa a qué grupo corresponde la región del mapa territorial en que se sitúa el punto cuyas coordenadas son precisamente las puntuaciones discriminantes citadas.* Por ejemplo, si las puntuaciones de la primera y segunda funciones discriminantes para un nuevo individuo son 4,5 y -5 respectivamente, este individuo se clasificará en el grupo 3, que es la zona del mapa territorial en la que cae el punto de coordenadas (4,5, -5).

La figura 16-45 muestra el diagrama de dispersión global para los cuatro grupos, que permite situar la posición de los casos y los centroides sobre las dos funciones discriminantes canónicas simultáneamente. Las coordenadas de cada caso serán sus puntuaciones discriminantes sobre las dos funciones. Como hay muchos casos, en la gráfica se han presentado también las posiciones de los centroides de grupo.

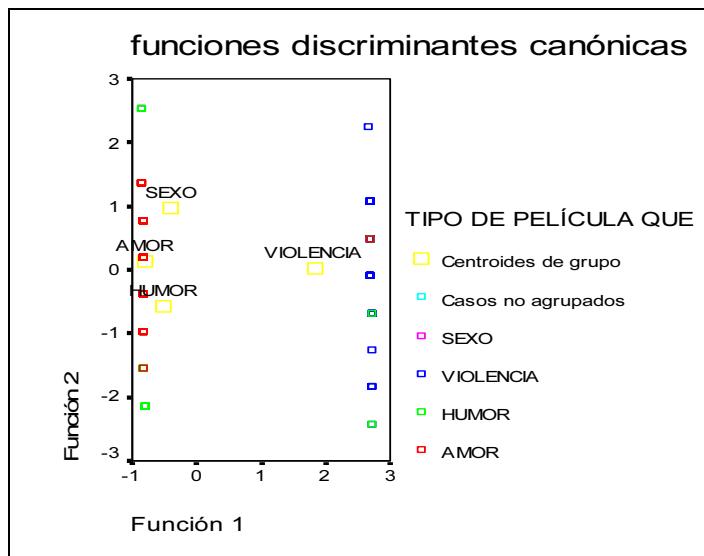


Figura 16-45

También es posible listar todos los casos con el grupo al que pertenecen, la probabilidad de pertenecer y la máxima probabilidad. Para ello usamos *Anализar → Informes → Resúmenes de casos* (figura 16-46) y rellenamos la pantalla de entrada como se indica en la figura 16-47. Al hacer clic en *Aceptar* se obtiene la tabla de resúmenes de casos de la figura 16-48.

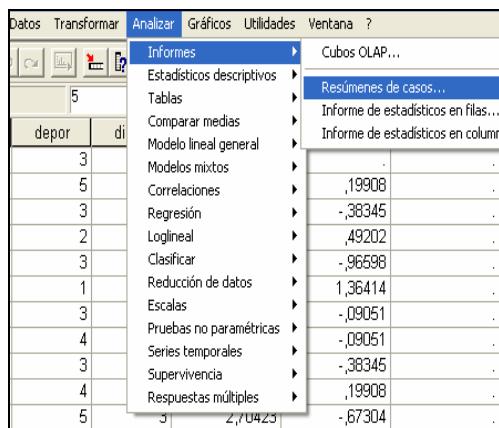


Figura 16-46

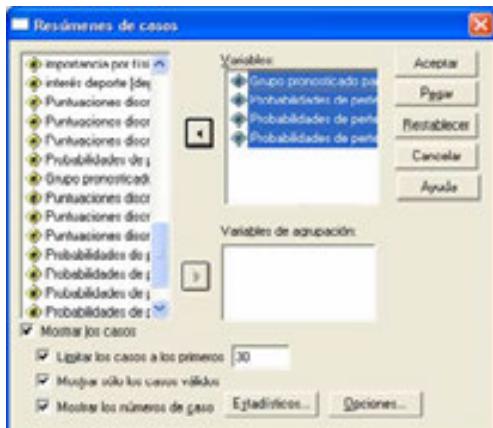


Figura 16-47

Resúmenes de casos(a)

	Número de caso	Grupo pronosticado para el análisis 1	Probabilidades de pertenencia al grupo 1 para el análisis 1	Probabilidades de pertenencia al grupo 2 para el análisis 1	Probabilidades de pertenencia al grupo 3 para el análisis 1
1	2	AMOR	,71597	,22245	,01153
2	3	AMOR	,65429	,30594	,01146
3	4	VIOLENCIA	,00603	,00402	,98784
4	5	AMOR	,57157	,40223	,01089
5	6	AMOR	,74997	,10287	,01020
6	7	VIOLENCIA	,00555	,00555	,98770
7	8	VIOLENCIA	,00555	,00555	,98770
8	9	AMOR	,65429	,30594	,01146
9	10	AMOR	,71597	,22245	,01153
10	11	VIOLENCIA	,00509	,00767	,98655
11	12	AMOR	,74997	,10287	,01020
12	13	AMOR	,71597	,22245	,01153
13	14	AMOR	,65429	,30594	,01146
14	15	AMOR	,71597	,22245	,01153
15	16	AMOR	,71597	,22245	,01153
16	17	AMOR	,71597	,22245	,01153
17	18	AMOR	,57157	,40223	,01089
18	19	AMOR	,74949	,15473	,01109
19	20	VIOLENCIA	,00656	,00290	,98684
20	21	VIOLENCIA	,00509	,00767	,98655
21	22	VIOLENCIA	,00603	,00402	,98784
22	23	VIOLENCIA	,00656	,00290	,98684
23	24	AMOR	,71597	,22245	,01153
24	25	AMOR	,65429	,30594	,01146
25	26	AMOR	,74949	,15473	,01109
26	27	HUMOR	,38070	,60682	,00859
27	28	AMOR	,71597	,22245	,01153
28	29	AMOR	,74949	,15473	,01109
29	30	VIOLENCIA	,00656	,00290	,98684
Total	N		29	29	29

a Limitado a los primeros 30 casos.

Figura 16-48

ANÁLISIS DE LA VARIANZA Y LA COVARIANZA

INTRODUCCIÓN AL ANÁLISIS ANOVA DE LA VARIANZA

El resultado de un experimento puede ser diferente al realizarlo varias veces, aunque las condiciones bajo las que se realiza sean siempre las mismas (aparentemente). Ello es consecuencia de las variaciones de muchos factores fuera de nuestro control que no permanecen exactamente constantes, y que influyen en el resultado del experimento. Si además se cambian las condiciones de realización variando los factores que influyen sustancialmente en el experimento, el resultado del mismo variará en mayor medida.

El análisis de la varianza descompone la variabilidad del resultado de un experimento en componentes independientes (variación total descompuesta en variaciones particulares).

Como ejemplo, podemos considerar los rendimientos de un mismo cultivo en parcelas diferentes, que aunque labradas en las mismas condiciones, producen cosechas que son distintas. Esta variabilidad de rendimientos es producida por multitud de factores controlables (abono, riego, etc.), donde cada factor puede presentar diferentes niveles (distintas cantidades o calidades de abonado, distintas intensidades de riego, etc.); también puede ser producida por otros factores no controlables (humedad relativa, clima, plagas, etc.).

Teóricamente, es posible dividir la variabilidad del resultado de un experimento en dos partes: la originada por los factores que influyen directamente en el resultado del experimento, estudiados en sus distintos niveles, y la producida por el resto de los factores con influencia en el resultado del experimento, variabilidad desconocida o no controlable, que se conoce con el nombre de *error experimental*.

El *análisis de la varianza simple* (ANOVA) se presenta cuando tenemos un solo factor (estudiado en sus distintos niveles) que influye sobre una *variable respuesta* que mide el resultado del experimento, y el resto de los factores forman el error experimental influyendo sobre la variable respuesta de forma no controlable. El factor se presenta con I niveles. Dentro de cada nivel analizamos una serie de observaciones del experimento en control (*unidades experimentales*) y su efecto sobre la variable respuesta; es decir, para cada nivel se repite el experimento varias veces (*replicación*).

Así, X_{ij} será la observación j -ésima de la variable respuesta relativa al i -ésimo nivel de factor (resultado obtenido para la variable respuesta en la repetición j -ésima del experimento para el i -ésimo nivel de factor).

En el ejemplo anterior, X_{ij} será el rendimiento obtenido (variable respuesta) bajo el nivel i de factor (abono) en la observación j -ésima (para cada nivel i de factor repetimos el cálculo del rendimiento n_i veces para recoger el efecto del error experimental).

Representamos por u_i la parte de X_{ij} debida a la acción del nivel i -ésimo de factor (en este caso un único factor).

Representamos por u_{ij} la variación causada por todos los factores no controlables (error experimental).

Podemos representar X_{ij} de la siguiente forma:

$$X_{ij} = u_i + u_{ij}$$

con i variando desde 1 hasta I , y j variando desde 1 hasta n_i . Se supone que u_{ij} es una variable normal de media 0 y desviación típica σ (varianza constante para todo i, j).

El hecho de que la media de u_{ij} sea cero, lleva a que la media de X_{ij} sea u_i . Por tanto, esta hipótesis exige que las n_i observaciones relativas al nivel i de factor tengan la misma media u_i .

El hecho de que la varianza de u_{ij} sea constante, nos dice que las perturbaciones tienen la misma variabilidad para los distintos niveles de factor, y que esa variabilidad es estable.

Otra de las hipótesis exigibles al modelo es que las medias de las variables u_{ij}, u_{rk} sean cero para todo i distinto de r , o para todo j distinto de k . Esta hipótesis implica la independencia de las observaciones X_{ij} .

Si hubiese dos factores conocidos e independientes que actuasen sobre la variable respuesta X_{ij} , el modelo de análisis de la varianza sería de *dos factores*, y al generalizar a múltiples factores tenemos el *modelo multifactorial de la varianza*.

Un modelo de análisis de la varianza es de *efectos fijos* cuando los resultados obtenidos sólo son válidos para esos determinados niveles de factor estudiados en el momento actual (factores constantes). Lo que ocurra a otros niveles de factor puede ser diferente.

Un modelo de análisis de la varianza es de *efectos aleatorios* cuando los resultados obtenidos son válidos sean cuales sean los niveles de factor empleados. Los factores son aleatorios (factores variables aleatorios), y los niveles estudiados de factor son una muestra de la población de niveles, que se supone infinita.

Un modelo es replicado si el experimento se repite varias veces para cada nivel de factor. En caso contrario, se dice que es un *modelo con una unidad por casilla*.

Un modelo de análisis de la varianza se llama *equilibrado* (balanced) cuando todos los n_i son iguales para cada nivel de factor i , o sea, cuando el número de observaciones para cada nivel de factor es siempre el mismo (el experimento se repite para cada nivel de factor el mismo número de veces). En caso contrario se llama *no equilibrado* (unbalanced).

ANÁLISIS DE LA VARIANZA SIMPLE (UN SOLO FACTOR): MODELO ANOVA UNIFACTORIAL DE EFECTOS FIJOS

En el apartado anterior hemos introducido los conceptos de efectos fijos y efectos aleatorios. En este apartado nos ocuparemos de los modelos de análisis de la varianza con un solo factor, distinguiendo entre efectos fijos y aleatorios.

Hemos visto que el modelo de análisis de la varianza para un solo factor se representa de la forma:

$$X_{ij} = u_i + u_{ij} \text{ ó } X_{ij} = u + A_i + u_{ij} \quad (u = \text{constante} \text{ y } A = \text{factor})$$

X_{ij} será el valor de la variable respuesta correspondiente a la observación j -ésima del i -ésimo nivel de factor (con i variando desde 1 hasta I , y j variando desde 1 hasta n_i). Los u_i (y los A_i) son constantes, y los u_{ij} son variables normales independientes de media 0 y desviación típica σ (varianza constante para todo i, j). Las X_{ij} son independientes y distribuidas $N(u_i, \sigma^2)$.

Podríamos representar las observaciones en un cuadro de la siguiente forma:

<i>Niveles</i>	<i>Observaciones</i>
1	$X_{11} X_{12} \dots X_{1j} \dots X_{1n_1}$
.
.
i	$X_{i1} X_{i2} \dots X_{ij} \dots X_{in_i}$
.
.
I	$X_{I1} X_{I2} \dots X_{Ij} \dots X_{In_I}$

El número total de elementos será $n = \sum_{i=1}^I n_i$.

El total del nivel i -ésimo será $X_{i..} = \sum_{j=1}^{n_i} X_{ij}$, y su media será $\bar{X}_{i..} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$.

El total general será $X_{..} = \sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^I X_{i..}$, y la gran media será:

$$\bar{X}_{..} = \frac{X_{..}}{n}$$

Representamos por u_i la parte de X_{ij} debida a la acción del nivel i -ésimo de factor (en este caso un único factor cualitativo).

Representamos por u_{ij} la variación en la variable respuesta causada por todos los factores no controlables (error experimental).

Se supone que u_{ij} es una variable normal de media 0 y desviación típica s (varianza constante para todo i,j).

Ya sabemos que el hecho de que la media de u_{ij} sea cero lleva a que la media de X_{ij} sea u_i . Por tanto, esta hipótesis exige que las n_i observaciones relativas al nivel i de factor tengan la misma media u_i . Además, el hecho de que la varianza de u_{ij} sea constante, nos dice que las perturbaciones tienen la misma variabilidad para los distintos niveles de factor, y que esa variabilidad es estable. También sabemos que otra de las hipótesis exigibles al modelo es que las medias de las variables u_{ij}, u_{rk} sean cero para todo i distinto de r , o para todo j distinto de k . Esta hipótesis implica la independencia de las observaciones X_{ij} .

Bajo estas hipótesis, podemos estimar el modelo, obteniendo como estimadores de los u_i los valores $\bar{X}_{i..}$, y como estimador de σ^2 el valor:

$$\hat{\sigma}^2 = \frac{1}{n-I} \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$$

La variación total viene dada por la cuasivarianza total, cuyo valor es:

$$ST^2 = \frac{1}{n-1} \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2$$

La variabilidad explicada o intervarianza (variación entre grupos, esto es, la variación entre los I grupos de observaciones de n_i unidades cada uno correspondientes a cada nivel de factor) viene dada por el valor:

$$S_b^2 = \frac{1}{I-1} \sum_{i=1}^I n_i (\bar{X}_{i\cdot} - \bar{X}_{..})^2$$

La variabilidad no explicada, o intravariación (variación dentro de grupos, esto es, la variación dentro de cada grupo de observaciones recogidas para los I niveles de factor en cada medición de la variable respuesta, es decir, en cada repetición del experimento), viene dada por el valor:

$$S_w^2 = \frac{1}{n-I} \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$$

Se demuestra que:

$$(n-1)ST^2 = (n-I)S_w^2 + (I-1)S_b^2$$

Es decir, que se puede desglosar la variación total de la muestra obtenida en las sucesivas repeticiones del experimento, considerada como un todo, en dos partes, dadas por la variación entre muestras y la variación dentro de las muestras.

Se demuestra que el cociente S_b^2/S_w^2 sigue una distribución F de Fisher con $I-1$ y $n-I$ grados de libertad. Este estadístico va a permitir contrastar la igualdad de medias para cada nivel de factor ($u_1 = u_2 = \dots = u_I = u$). De esta forma, se contrastará si los distintos niveles de factor influyen o no significativamente en la misma medida en la variable respuesta.

Por otra parte, también se demuestra que el estadístico $\frac{I-1}{\sigma^2} S_b^2$ sigue una chi-cuadrado con $I-1$ grados de libertad.

También se sabe que el estadístico $\frac{N-I}{\sigma^2} S_w^2$ sigue una chi-cuadrado con $I - 1$ grados de libertad.

Estos dos estadísticos permiten calcular intervalos de confianza para σ .

S_w^2 es un estimador insesgado para σ^2 .

S_b^2 sólo es estimador insesgado para σ bajo la hipótesis de que los u_i son todos iguales ($u_i = u$ para todo i).

A S_b^2 se le suele llamar *cuadrado medio del factor*: $CM(A)$.

A S_w^2 se le suele llamar *cuadrado medio del error*: $CM(E)$.

Con lo que la F de Fisher es el cociente entre el cuadrado medio del factor y el cuadrado medio del error.

$$F = CM(A)/CM(E)$$

Toda esta información suele resumirse en una tabla, llamada tabla ANOVA, de la siguiente forma:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Entre grupos	$(I-1)S_b^2 = SC(A)$	$I-1$	$S_b^2 = CM(A)$	$CM(A)$
Intra-grupos	$(n-I)S_w^2 = SC(E)$	$n-I$	$S_w^2 = CM(E)$	$CM(E)$
Total (corregida)	$(n-1)ST = SC(T)$	$n-1$	S^2	

Una vez estimado el modelo, será necesario comprobar mediante distintos tests si las hipótesis básicas del mismo no están en contradicción con los datos observados.

Para contrastar la normalidad de los errores experimentales u_{ij} (hipótesis de normalidad), suele usarse el test W de SHAPIRO y WILK, un contraste de la chi-cuadrado o el test de KOLMOGOROV-SMIRNOV..

Para contrastar la igualdad de varianzas de los u_{ij} (hipótesis de homoscedasticidad), suele usarse el test de BARLETT, el test Q de COCHRAN y el test de HARTLEY.

Para contrastar la independencia de las observaciones, o no correlación de los residuos (hipótesis de no autocorrelación), suelen utilizarse el coeficiente de correlación serial o el test de rachas.

Ya sabemos que es interesante contrastar si es aceptable la hipótesis de que las medias de todos los grupos de observaciones obtenidas al repetir el experimento para cada nivel de factor son idénticas ($\mu_1 = \mu_2 = \dots = \mu_I = \mu$). Si los contrastes diesen como resultado que esta hipótesis es cierta, la pertenencia a un grupo o a otro sería irrelevante, y podríamos considerar todas las observaciones como una muestra de una única población. Un enfoque alternativo de esta hipótesis, que conduce al mismo resultado, es considerar los grupos idénticos si las diferencias entre sus medias son pequeñas.

Se pueden construir intervalos de confianza para las diferencias entre medias de distintos grupos ($\mu_i - \mu_j$), con el fin de estimar si existen diferencias entre ellos. Como norma general, si el intervalo contiene al cero se suele aceptar la hipótesis de medias iguales para los grupos. También se pueden construir intervalos de confianza para la varianza del error experimental y para cocientes de varianzas.

En general, cuando estudiamos mediante el análisis de la varianza el comportamiento de los niveles de un factor fijo, no se persigue como única finalidad del análisis saber si globalmente los distintos niveles de factor son significativamente distintos entre sí en su efecto sobre la variable respuesta (aspecto de un evidente interés), sino que, lógicamente, estaremos interesados en conocer, una vez contrastado que las diferencias son significativas, qué niveles producen un efecto superior al de otros sobre la variable respuesta. Para ello existen diferentes contrastes que efectúan comparaciones múltiples entre las I medias o combinaciones lineales de ellas (test de recorrido múltiple de DUNCAN, test de BONFERRONI, test SNK de STUDENT_NEWMAN_KEULS, test HSD de TUKEY, test de la diferencia mínima significativa, test de SCHEFFE y test de TUKEY de comparaciones múltiples).

MODELO UNIFACTORIAL DE EFECTOS ALEATORIOS

Ya sabemos que estamos ante un modelo unifactorial de efectos aleatorios, también llamado *modelo de componentes de la varianza*, cuando de la población total de niveles del factor (supuesta infinita o suficientemente grande como para considerarla infinita), los I niveles del factor que se utilizan en el experimento se han elegido aleatoriamente. En este caso, el modelo ANOVA de efectos aleatorios se formulará de la siguiente forma:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Una formulación equivalente, si consideramos que $\mu_i = \mu + \beta_i$, sería:

$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$$

donde:

- μ es una constante.
- β_i , para $i = 1, \dots, I$, son variables aleatorias independientemente distribuidas $N(0, \sigma^2_\beta)$.
- ε_{ij} , para $i = 1, \dots, I$ y $j = 1, \dots, n_i$, son v.a.i.i.d. $N(0, \sigma^2)$.
- β_i y ε_{ij} , para $i = 1, \dots, I$ y $j = 1, \dots, n_i$, son variables aleatorias independientes. (v.a.i.i.d.=variables aleatorias independientes idénticamente distribuidas)

Para este modelo se verifica que $E[Y_{ij}] = \mu$, y la varianza de Y_{ij} , denotada por σ_Y^2 , será $V[Y_{ij}] = \sigma_Y^2 = \sigma^2_\beta + \sigma^2$, donde σ^2_β y σ^2 se llaman *componentes de la varianza*, razón por la cual el modelo de efectos aleatorios se denomina modelo de componentes de la varianza. En este modelo, los cálculos del análisis de la varianza para la suma de cuadrados son idénticos a los del modelo de efectos fijos, de manera que la descomposición de suma de cuadrados sigue siendo válida, así como la descomposición de los grados de libertad. La definición de los cuadrados medios es también exactamente la misma, al igual que el contraste de la F para la igualdad de medias. Sin embargo, como el factor es una variable aleatoria, habrá que considerar su varianza σ^2_β .

Aparece como novedad las estimaciones de σ^2_β y σ^2 , que vienen dadas mediante las expresiones siguientes:

$$\hat{\sigma}_\beta^2 = (CM(A) - CM(E))/n_0 \quad y \quad \hat{\sigma}^2 = CM(E)$$

Para un modelo equilibrado se tiene que $n_0 = n_1 = n_2 = \dots = n_I$, y para un modelo no equilibrado se tiene que:

$$n_0 = \frac{\left[\sum_{i=1}^I n_i - \frac{\sum_{i=1}^I n_i^2}{\sum_{i=1}^I n_i} \right]}{I - 1}$$

También aparece como novedad en el modelo de efectos aleatorios el contraste:

$$\begin{aligned} H_0: \sigma^2_\beta &= 0 \\ H_1: \sigma^2_\beta &> 0 \end{aligned}$$

Para llevar a cabo este contraste, se emplea el estadístico $F = CM(A)/CM(E)$, que se distribuye, bajo H_0 , según una ley de Fisher-Snedecor, con $I - 1$ y $n_0 - I$ grados de libertad. En consecuencia, el contraste será:

Aceptar $H_0: \sigma^2_{\beta} = 0$, cuando $F^* \leq F_{\alpha; I-1, n_o - I}$
 Aceptar $H_1: \sigma^2_{\beta} > 0$, cuando $F^* > F_{\alpha; I-1, n_o - I}$

Los intervalos de confianza para la media de un solo tratamiento y para la diferencia entre las medias de un par de tratamientos basados en los datos obtenidos utilizando un diseño completamente aleatorizado son similares a los intervalos de confianza habituales. El intervalo de confianza para la media del tratamiento i es:

$$\bar{T}_i \pm \frac{t_{\alpha/2}S}{\sqrt{n_i}} = \frac{\bar{x}_{i\cdot}}{n_i} \pm \frac{t_{\alpha/2}S}{\sqrt{n_i}} = \bar{x}_{i\cdot} \pm \frac{t_{\alpha/2}S}{\sqrt{n_i}}$$

El intervalo de confianza para la diferencia entre los tratamientos i y j ($A_i - A_j$) será:

$$\bar{T}_i - \bar{T}_j \pm t_{\alpha/2}S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \frac{\bar{x}_{i\cdot}}{n_i} - \frac{\bar{x}_{j\cdot}}{n_j} \pm t_{\alpha/2}S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \bar{x}_{i\cdot} - \bar{x}_{j\cdot} \pm t_{\alpha/2}S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$S = \sqrt{CM(E)} = \sqrt{\frac{SC(E)}{n - I}} = \sqrt{\frac{SC(E)}{n_1 + n_2 + \dots + n_I - I}}$$

En general, un intervalo de confianza para la combinación lineal de tratamientos $\sum c_i A_i$ puede calcularse como sigue:

$$\sum_i c_i \bar{T}_i \pm t_{\alpha/2}S \sqrt{\sum_i \frac{c_i^2}{n_i}} = \sum_i c_i \frac{\bar{x}_{i\cdot}}{n_i} \pm t_{\alpha/2}S \sqrt{\sum_i \frac{c_i^2}{n_i}} = \sum_i c_i \bar{x}_{i\cdot} \pm t_{\alpha/2}S \sqrt{\sum_i \frac{c_i^2}{n_i}}$$

ANÁLISIS DE LA VARIANZA CON VARIOS FACTORES: MODELO BIFACTORIAL DE EFECTOS FIJOS ANOVA IIF

El análisis multifactorial de la varianza se presenta cuando dos o más factores (variables independientes) afectan a la variable respuesta (variable dependiente). Para cada factor tendremos varios niveles, que dividen la población total en grupos de tratamiento. En la terminología del diseño de experimentos, suelen denominarse *tratamientos* los distintos niveles de cada factor.

Un concepto importante a tener en cuenta en el modelo multifactorial de la varianza es el análisis de la *interacción* entre las variables (factores). Se dice que hay interacción cuando una variable independiente A afecta a la variable dependiente de manera distinta según los diferentes niveles de otra variable independiente B . En síntesis, se trata, por tanto, de analizar si las variables independientes (factores) producen efectos distintos en función de los niveles de las otras variables independientes.

Para un modelo factorial de dos factores A y B de efectos fijos tendremos la expresión general:

$$X_{ijk} = u + A_i + B_j + AB_{ij} + E_{ijk} \quad i=1..t, j=1..r, k=1..n_{ij} \quad (u = \text{constante})$$

Los términos A_i y B_j representan los efectos de los factores A y B (*efectos principales*), y son constantes sujetas a las restricciones:

$$\sum_{i=1}^t A_i = \sum_{j=1}^r B_j = 0$$

Los términos AB_{ij} representa el *efecto de la interacción* entre los factores A y B , y son constantes sujetas a las restricciones:

$$\sum_{i=1}^t (AB)_{ij} = \sum_{j=1}^r (AB)_{ij} = 0$$

El término E_{ijk} representa el error experimental, que corresponderá a una variable aleatoria normal de media cero y varianza σ^2 constante para cada k (las variables E_{ijk} han de ser independientes).

Los datos podrían representarse en una tabla como la siguiente:

<i>Factor B →</i> <i>(niveles)</i>	1	2	<i>r</i>
<i>Factor A</i> ↓				
1	X_{111} X_{112} .	X_{121} X_{122}	X_{1r1} X_{1r2} .
	X_{11n1}	X_{12n12}		X_{1m1r}
2	X_{211} X_{212} .	X_{221} X_{222}	X_{2r1} X_{2r2} .
	X_{21n21}	X_{22n22}		X_{2m2r}
<i>t</i>	X_{t11} X_{t12} .	X_{t21} X_{t22}	X_{tr1} X_{tr2} .
	X_{t1n1}	X_{t2n2}	...	X_{tmtr}

La descomposición de la suma de cuadrados total *para el caso en que n_{ij} es constante ($n_{ij}=s \forall i,j$)*, puede hacerse ahora de la forma:

$SCT = SCA + SCB + SCAB + SCE$, donde:

$$SCT = \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (X_{ijk} - \bar{X}...)^2 = \text{suma de cuadrados total.}$$

$$SCA = \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (\bar{X}_{i..} - \bar{X}...)^2 = \text{suma de cuadrados del factor } A.$$

$$SCB = \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (\bar{X}_{.j.} - \bar{X}...)^2 = \text{suma de cuadrados del factor } B.$$

$$SCAB = \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}...)^2 = \text{suma de la interacción.}$$

$$SCE = \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (X_{ijk} - \bar{X}_{ij.})^2 = \text{suma de cuadrados del error.}$$

También se usa la suma de cuadrados entre tratamientos, cuyo valor es:

$$SCTR = \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s (\bar{X}_{ij.} - \bar{X}...)^2$$

$$\bar{X}... = \frac{1}{trs} \sum_{i=1}^t \sum_{j=1}^r \sum_{k=1}^s X_{ijk} = \text{gran media.}$$

$$\bar{X}_{i..} = \frac{1}{rs} \sum_{j=1}^r \sum_{k=1}^s X_{ijk} = \text{media de la fila } i.$$

$$\bar{X}_{.j.} = \frac{1}{ts} \sum_{i=1}^t \sum_{k=1}^s X_{ijk} = \text{media de la columna } j.$$

$$\bar{X}_{ij.} = \frac{1}{s} \sum_{k=1}^s X_{ijk} = \text{media de la clase } (i,j) \text{ para } i = 1 \dots t, j = 1 \dots r.$$

Para el caso en que n_{ij} es constante ($n_{ij} = s$), la información del modelo se resume en el siguiente cuadro del análisis de la varianza (Tabla ANOVA II):

Fuente de variación	Suma cuadrados	Grados de libertad	Cuadrados medios	F
Entre tratamientos	SCTr	tr-1	CMT _r =SCT _r /(tr-1)	
Entre filas (factor A)	SCA	t-1	CMA=SCA/(t-1)	CMA/CME
Entre colum .(factor B)	SCB	r-1	CMB=SCB/(r-1)	CMB/CME
Interacciones (AB)	SCAB	(t-1)(r-1)	CMAB=SCAB/(t-1)(r-1)	CMAB/CME
Error	SCE	tr(s-1)	CME=SCE/tr(s-1)	
Total	SCT	trs-1		

La última columna de la Tabla ANOVA II expresa los estadísticos a utilizar en los contrastes de la F de Fisher Snedecor.

Lo primero que interesa conocer cuando se estudian dos factores es si se puede aceptar que los efectos medios de interacción, $(AB)_{ij}$, son iguales. Este contraste se formulará de la siguiente forma:

$$H_0(AB): (AB)_{11} = (AB)_{12} = \dots = (AB)_{tr} = 0$$

$$H_1(AB): \text{no todos los } (AB)_{ij} \text{ son iguales}$$

Si $H_0(AB)$ es cierta, el estadístico $F^{***} = CMAB / CME$ sigue una distribución de probabilidad F de Fisher Snedecor con $(t-1).(r-1)$ y $tr(s-1)$ grados de libertad, por tratarse del cociente de distribuciones chi-cuadrado independientes, estando cada una dividida por sus grados de libertad. Por tanto, fijado un nivel de significación α , la regla de decisión del contraste de igualdad de medias de la interacción AB será:

Aceptar $H_0(AB)$ cuando $F^{***} \leq F_{\alpha;(t-1),(r-1),tr,(s-1)}$

Aceptar $H_1(AB)$ cuando $F^{***} > F_{\alpha;(t-1),(r-1),tr,(s-1)}$

Para contrastar la igualdad de los niveles medios del factor A , se tiene que el estadístico $F^* = CMA / CME$ se distribuye según una ley F de Fisher Snedecor con $t-1$ y $t.r.(s-1)$ grados de libertad. Por tanto, fijado un nivel de significación α , la regla de decisión del contraste de igualdad de medias del factor A será:

Aceptar $H_0(A)$ cuando $F^* \leq F_{\alpha;t-1,t.r.(s-1)}$

Aceptar $H_1(A)$ cuando $F^* > F_{\alpha;t-1,t.r.(s-1)}$

De manera análoga, se tiene que el estadístico del contraste de igualdad de medias del factor B será $F^{**} = CMB / CME$, cuya distribución es una ley F de Fisher Snedecor con $r-1$ y $t.r.(s-1)$ grados de libertad. La regla de decisión, a un nivel de significación α , será:

Aceptar $H_0(B)$ cuando $F^{**} \leq F_{\alpha;r-1,t.r.(s-1)}$

Aceptar $H_1(B)$ cuando $F^{**} > F_{\alpha;r-1,t.r.(s-1)}$

En cuanto a la estimación de la varianza en el modelo ANOVA II, se observa que la estimación de la varianza poblacional común σ^2 resulta ser el cuadrado medio del error $\hat{\sigma}^2 = CME$. Este estimador es un estimador insesgado de la varianza poblacional, siempre que se pueda garantizar que las t^*r^*s poblaciones consideradas tienen la misma varianza.

En cuanto a la estimación de los parámetros del modelo ANOVA II, tenemos los siguientes resultados:

$$\hat{\sigma}^2 = CME, \hat{u} = \bar{Y} \dots, \hat{A}_i = \bar{Y}_{i..} - \bar{Y} \dots, \hat{B}_j = \bar{Y}_{.j..} - \bar{Y} \dots; \text{ y } \hat{AB}_{ij} = \bar{Y}_{ij..} - \bar{Y}_{i..} - \bar{Y}_{.j..} + \bar{Y} \dots$$

El desarrollo del modelo ANOVA II general se ha realizado bajo la hipótesis del mismo número de observaciones, s , para cada tratamiento o combinación de niveles de los factores. Cuando no se cumple dicha hipótesis, el análisis de la varianza para un estudio de dos factores se hace más complejo, y ya no son válidas las fórmulas de descomposición de suma de cuadrados, aunque se mantiene la filosofía de la descomposición de la suma de cuadrados total y de los grados de libertad. A partir de las nuevas expresiones de sumas de cuadrados, se consideran los cuadrados medios y el cociente correspondiente entre cuadrados medios, de manera análoga. Una forma sencilla de abordar este problema y obtener las sumas de cuadrados apropiadas para realizar los contrastes de hipótesis sobre los efectos de interacción de los factores y sobre los efectos principales de los factores, es considerar el análisis de la varianza desde la perspectiva del análisis de regresión.

Los intervalos de confianza para combinaciones lineales de tratamientos se llevan a cabo de la misma forma que en el modelo de un solo factor, pero teniendo en cuenta que si las interacciones son significativamente distintas de cero, se utilizará CMAAB en lugar de CME en las fórmulas.

En general, un intervalo de confianza para la combinación lineal de tratamientos $\sum c_i A_i$ cuando las interacciones son significativas puede calcularse como sigue:

$$\sum_i c_i \bar{x}_{i..} \pm t_{\alpha/2} \sqrt{\frac{CMAB \sum_i c_i^2}{rs}}$$

MODELO BIFACTORIAL GENERAL CON EFECTOS ALEATORIOS ANOVA IIA

Hasta ahora, en el modelo ANOVA II se suponía que existían unos niveles fijos para cada uno de los factores. Por esa razón, al modelo se le llama modelo ANOVA II general con efectos fijos, abreviadamente modelo ANOVA IIF general.

Supongamos ahora que cada uno de los dos conjuntos de niveles de los factores se puede considerar una muestra de una población suficientemente grande sobre la que se van a realizar estudios. En este caso, se dice que estamos en presencia del *modelo ANOVA II general con efectos aleatorios*, abreviadamente modelo ANOVA IIA general.

Por tanto, el modelo ANOVA II general con efectos aleatorios se formula de la siguiente forma:

$$Y_{ijl} = \mu + \beta_i + \delta_j + (\beta\delta)_{ij} + \varepsilon_{ijl}$$

donde, para $i = 1, \dots, h; j = 1, \dots, k; l = 1, \dots, t$; se verifica que μ es una constante, β_i son v.a. independientes distribuidas $N(0, \sigma_\beta^2)$, δ_j son v.a. independientes distribuidas $N(0, \sigma_\delta^2)$, $(\beta\delta)_{ij}$ son v.a. independientes distribuidas $N(0, \sigma_{\beta\delta}^2)$, ε_{ijl} son v.a. independientes distribuidas $N(0, \sigma^2)$, y $\beta_i, \delta_j, (\beta\delta)_{ij}$ y ε_{ijl} son v.a. independientes dos a dos.

Para este modelo se verifica que $E [Y_{ijl}] = \mu$, y la varianza de Y_{ijl} , notada por σ_y^2 , viene dada por $V [Y_{ijl}] = \sigma_y^2 = \sigma_\beta^2 + \sigma_\delta^2 + \sigma_{\beta\delta}^2 + \sigma^2$.

En este modelo, los cálculos del análisis de la varianza para las sumas de cuadrados son idénticos a los realizados en el modelo de efectos fijos. De manera análoga, los grados de libertad y los cuadrados medios son exactamente los mismos. Los modelos ANOVA IIA general y ANOVA IIM general (modelo general con efectos mixtos) difieren del modelo ANOVA IIF general en los cuadrados medios esperados y en la elección de los estadísticos de los contrastes.

Como es habitual, para contrastar los efectos de los factores en el modelo ANOVA IIA se construye un estadístico comparando dos cuadrados medios, que, bajo H_0 , sigue una distribución F de Snedecor. A partir de ello se construye la regla de decisión en la forma habitual.

El contraste sobre la presencia de efectos del factor A vendrá formulado por:

$$\begin{aligned} H_0: \sigma_\beta^2 &= 0 \\ H_1: \sigma_\beta^2 &> 0 \end{aligned}$$

El estadístico del contraste será $F^* = CMA / CMAB$, que se distribuye, bajo H_0 , según una ley de Snedecor con $h-1$ y $(h-1).(k-1)$ grados de libertad. En consecuencia, el contraste será:

Aceptar $H_0: \sigma_\beta^2 = 0$, cuando $F^* \leq F_{\alpha;h-1,(h-1),(k-1)}$

Aceptar $H_1: \sigma_\beta^2 > 0$, cuando $F^* > F_{\alpha;h-1,(h-1),(k-1)}$

De manera análoga, el contraste sobre la presencia de efectos del factor B , se realizará mediante el estadístico $F^{**} = CMB / CMAB$ que se distribuye, bajo H_0 , según una ley de Snedecor con $k-1$ y $(h-1).(k-1)$ grados de libertad. En consecuencia, el contraste será:

Aceptar $H_0: \sigma_\delta^2 = 0$, cuando $F^{**} \leq F_{\alpha;k-1,(h-1),(k-1)}$

Aceptar $H_1: \sigma_\delta^2 > 0$, cuando $F^{**} > F_{\alpha;k-1,(h-1),(k-1)}$

Por último, el contraste sobre la presencia de efectos de interacción entre el factor A y el B utilizará el estadístico $F^{***} = CMAB / CME$, que se distribuye, bajo H_0 , según una ley de Snedecor con $(h-1).(k-1)$ y $h.k.(t-1)$ grados de libertad. En consecuencia, el contraste será:

Aceptar $H_0: \sigma_{\beta\delta}^2 = 0$, cuando $F^{***} \leq F_{\alpha;(h-1),(k-1).h.k.(t-1)}$

Aceptar $H_1: \sigma_{\beta\delta}^2 > 0$, cuando $F^{***} > F_{\alpha;(h-1),(k-1).h.k.(t-1)}$

Para estimar las componentes de la varianza, σ^2 , σ_β^2 , σ_δ^2 y $\sigma_{\beta\delta}^2$, se utilizan los estimadores CME , $(CMA - CMAB)/k.t$, $(CMB - CMAB)/h.t$ y $(CMAB - CME)/t$ respectivamente.

MODELO BIFACTORIAL GENERAL CON EFECTOS MIXTOS ANOVA IIM

En un modelo factorial de dos factores, pueden ser los dos fijos, los dos aleatorios, o uno aleatorio y otro fijo. En este último caso, estamos ante un *modelo mixto*. Cuando en uno de los dos factores se consideran niveles fijos y en el otro niveles aleatorios, se dirá que se trata de un modelo ANOVA II general con efectos mixtos, abreviadamente modelo ANOVA IIM general. Si el factor A tiene niveles fijos y el factor B tiene niveles aleatorios, los efectos β_i son constantes, los efectos δ_j son v.a., y los efectos de interacción $(\beta\delta)_{ij}$ también son v.a. al serlo los δ_j . Suponiendo tamaños muestrales iguales para cada tratamiento, se tiene que el modelo ANOVA IIM general con efectos mixtos se formula de la siguiente forma:

$$Y_{ijl} = \mu + \beta_i + \delta_j + (\beta\delta)_{ij} + \varepsilon_{ijl}$$

donde se verifica que μ es una constante, β_i son constantes tales que $\sum \beta_i = 0$, δ_j para $j = 1, \dots, k$ son v.a. independientes distribuidas $N(0, \sigma_{\beta\delta}^2)$, $(\beta\delta)_{ij}$ para $i = 1, \dots, h$, son v.a.i.i.d. $N(0, \sigma_{\beta\delta}^2(h-1/h))$ sujetas a las restricciones $\sum (\beta\delta)_{ij} = 0$ para todo $j = 1, \dots, k$, ε_{ijl} para $i=1, \dots, h, j = 1, \dots, k$ y $l = 1, \dots, t$ son v.a. independientes distribuidas $N(0, \sigma^2)$, y δ_j , $(\beta\delta)_{ijl}$ y ε_{ijl} son v.a. independientes dos a dos.

El contraste sobre efectos del factor fijo A vendrá formulado por:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_h = 0$$

H_1 : No todos los β_i son iguales

El estadístico del contraste será $F^* = CMA / CMAB$, a diferencia del modelo ANOVA IIF donde se empleaba CME en el denominador. La razón está en que son, en el modelo ANOVA IIM general, CMA y $CMAB$ los que tienen el mismo valor esperado bajo H_0 . Dicho estadístico se distribuye según una ley F de Fisher Snedecor con $h-1$ y $(h-1).(k-1)$ grados de libertad. En consecuencia, el contraste será:

Aceptar H_0 , cuando $F^* \leq F_{\alpha; h-1, (h-1).(k-1)}$

Aceptar H_1 , cuando $F^* > F_{\alpha; h-1, (h-1).(k-1)}$

El contraste sobre la presencia de efectos del factor aleatorio B , se realizará mediante el estadístico $F^{**} = CMB / CME$, a diferencia del modelo ANOVA IIA, donde se empleaba $CMAB$ en el denominador. La razón está en que, en el modelo ANOVA IIM general, son CMB y CME los que tienen el mismo valor esperado bajo H_0 . Dicho estadístico se distribuye según una ley F de Fisher Snedecor con $k-1$ y $h.k.(t-1)$ grados de libertad. En consecuencia, el contraste será:

Aceptar $H_0: \sigma_\delta^2 = 0$, cuando $F^{**} \leq F_{\alpha; (k-1), h.k.(t-1)}$

Aceptar $H_1: \sigma_\delta^2 > 0$, cuando $F^{**} > F_{\alpha; (k-1), h.k.(t-1)}$

Para el contraste sobre la presencia de efectos de interacción entre el factor A y el B , es válido lo analizado en el modelo ANOVA IIA.

En este modelo mixto es posible estimar los efectos del factor fijo mediante $\hat{u} = \bar{Y}_{...}$ y $\hat{\beta}_i = \bar{Y}_{i..} - \bar{Y}_{...}$, y para estimar las componentes de la varianza σ^2 , σ_δ^2 y $\sigma_{\beta\delta}^2$, se utilizan los estadísticos CME , $(CMB - CME)/h.t$ y $(CMAB - CME)/t$ respectivamente.

MODELO EN BLOQUES ALEATORIZADOS

En los análisis de la varianza realizado, hasta ahora hemos supuesto que las unidades experimentales se asignan a los niveles de los factores o tratamientos (o viceversa) completamente al azar en la realización del experimento, dando lugar así a los *diseños completamente aleatorizados*. La situación idónea es aquélla en la que existe una gran homogeneidad entre las unidades experimentales (dos unidades experimentales sometidas a distintos tratamientos presentan leves diferencias), pero cuando las unidades experimentales no son homogéneas, se recomienda utilizar el *diseño de bloques al azar*. En este caso, los tratamientos se aplican a grupos homogéneos de unidades experimentales (ya no hay una asignación aleatoria de las unidades experimentales a los tratamientos, o viceversa).

Cada bloque homogéneo equivale a un grupo experimental, considerándose dichos bloques equivalentes a los niveles de los factores, si bien en este diseño los bloques no se consideran como factores a estudiar, sino simplemente como una forma de controlar la varianza intra grupo (error experimental). La asignación de unidades experimentales a cada bloque se realiza de forma aleatoria. *El diseño de bloques al azar es completo* si todos los bloques tienen representación en todos los tratamientos.

El modelo de bloques al azar corresponde a un modelo ANOVA IIF de dos factores fijos sin interacciones, en el que uno de los factores corresponde a los bloques. Estos modelos se tratan como el modelo de dos factores fijos sin interacción.

Si consideramos b bloques y k tratamientos podemos representarlos como sigue:

Tratamiento	Bloque 1	Bloque 2	...	Bloque b	Total	Media
1	y_{11}	y_{12}	...	y_{1b}	$T_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	...	y_{2b}	$T_{2\cdot}$	$\bar{y}_{2\cdot}$
.
k	y_{k1}	y_{k2}	...	y_{kb}	$T_{k\cdot}$	$\bar{y}_{k\cdot}$
Total	$T_{\cdot 1}$	$T_{\cdot 2}$...	$T_{\cdot b}$	$T_{..}$	
Media	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$...	$\bar{y}_{\cdot b}$		$\bar{y}_{..}$

El cuadro del análisis de la varianza para el diseño de bloques completamente aleatorizados (ANOVA IIF de dos factores fijos sin interacciones) quedará como sigue:

Fuente de variación	Suma cuadrados	Grados de libertad	Cuadrados medios	F
Entre filas (factor A)	SCA	$k-1$	$CMA=SCA/(k-1)$	CMA/CME
Entre colum.(bloques)	SCB	$b-1$	$CMB=SCB/(b-1)$	
Error	SCE	$(k-1)(b-1)$	$CME=SCE/(k-1)(b-1)$	
Total	SCT	$bk-1$		

Un *diseño de medidas repetidas* (o *diseño intrasujeto*) es un caso particular de diseño de bloques al azar, que consiste en que cada bloque está formado por un solo individuo al que se aplican todos los tratamientos.

En este tipo de diseño se puede dar el *efecto de superposición (carry-over effect)*, que se produce cuando se administra un tratamiento antes de que haya terminado el efecto de un tratamiento anterior. Este efecto puede controlarse aumentando el tiempo entre los tratamientos. En el diseño en medidas repetidas puede darse también el *efecto de aprendizaje*, que se produce cuando la simple repetición mejora la respuesta, independientemente de cualquier tratamiento. También puede darse en el diseño de medidas repetidas el *efecto latencia*, que se produce cuando un tratamiento activa el efecto de un tratamiento anterior que permanecía en estado de latencia.

El *diseño split-splot o diseño en parcelas divididas*, es una extensión del diseño de bloques al azar, cuyo origen es el análisis agrario. El concepto *split-splot* se refiere a una parcela de terreno que se subdivide (*split*) en varias porciones (*splot*). Estos diseños también se aplican a la investigación educativa. Este modelo se utiliza cuando se combinan dos factores A y B y se obtienen réplicas organizadas en bloques. El factor bloque C tiene un efecto principal, pero no interacciona con A y B a la vez. En este diseño se comparan a tratamientos (factor A), que se asignan aleatoriamente en b bloques o parcelas (factor B), a razón de a tratamientos por bloque. Se divide cada una de las ab parcelas, y se asignan al azar c subtratamientos a estas divisiones (factor C). Se supone que actúan los efectos principales A , B , C , la interacción AC y la interacción AB . La interacción entre A y los bloques es debida a que éstos no pueden considerarse completamente homogéneos. Sin embargo, se supone que cada una de las ab parcelas dentro de los bloques son homogéneas para que los subtratamientos C no interaccionen con los bloques.

MODELO ANOVA FACTORIAL CON TRES FACTORES

Para un modelo factorial de tres factores A , B y C , tendríamos la expresión general:

$$X_{ijkl} = u + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} * ABC_{ijk} + E_{ijkl} \quad i=1..t, j=1..r, k=1..s, l=1..n_{ijk}$$

Los términos A_i , B_j y C_k representan los efectos de los factores A , B y C (efectos principales). El término AB_{ij} representa el efecto de la interacción entre los factores A y B . El término AC_{ik} representa el efecto de la interacción entre los factores A y C . El término BC_{jk} representa el efecto de la interacción entre los factores B y C . El término ABC_{ijk} representa la interacción triple entre los factores A , B y C . El término E_{ijkl} representa el error experimental, que corresponderá a una variable aleatoria normal de media cero y varianza constante para cada l . Las variables E_{ijkl} han de ser independientes. El modelo también puede considerarse con término constante.

En un modelo factorial de tres factores, pueden ser los tres fijos, los tres aleatorios, uno aleatorio y dos fijos, o dos aleatorios y el otro fijo. En un modelo multifactorial los niveles de cada factor (tratamientos) suelen estar combinados con todos los niveles de los restantes factores. En el caso de que ciertos niveles de determinados factores estén ligados solamente a ciertos niveles de otros, estamos ante un *diseño jerárquico*.

En un diseño jerárquico, los niveles de cada factor están incluidos en los niveles de otro factor, estableciéndose así una jerarquía de dependencias entre los distintos niveles de los diferentes factores. Un modelo jerárquico es *anidado* cuando cada nivel de un factor se corresponde sólo con un nivel de otro factor. En este tipo de modelos no existen interacciones, ya que esto sólo es posible cuando todos los niveles de un determinado factor se cruzan con todos los niveles de los demás factores. Un modelo jerárquico es *cruzado* cuando todos los niveles de un factor aparecen en todos los niveles de resto de los factores.

MODELO EN CUADRADO LATINO

El *diseño en cuadrado latino* es aquél en el que se tienen en cuenta tres factores, con la particularidad de que los tres factores tienen el mismo número de niveles. Dos de los factores operan como bloques con el fin de reducir los errores experimentales. El objetivo del análisis es determinar si el tercer factor (factor de tratamiento) tiene o no una influencia significativa sobre la variable dependiente. En un diseño en cuadrado latino hay una sola observación por casilla, por lo que en ningún caso es posible tratar efectos de interacción. En el diseño en cuadrado latino existen dos factores de bloque, a diferencia del diseño de bloques al azar, en el que sólo hay un factor de bloque. Otra característica diferenciadora del diseño en cuadrado latino es que los dos factores de bloque, así como el factor de tratamiento, tienen el mismo número de grupos o niveles.

El modelo que puede establecerse en este caso es el siguiente:

$$Y_{ij(k)} = \mu + \alpha_i + \beta_j + \gamma_k + u_{ij(k)} \quad i,j,k = 1,\dots,r$$

Considerando la restricción $\sum_{i=1}^r \alpha_i = \sum_{j=1}^r \beta_j = \sum_{k=1}^r \gamma_k = 0$, la tabla del análisis de la varianza para un cuadrado latino $r \times r$ quedará como sigue:

Fuente de variación	Suma cuadrados	Grados de libertad	Cuadrados medios	F
Entre filas (renglones)	SCA	$r-1$	$CMA=SCA/(r-1)$	
Entre columnas	SCB	$r-1$	$CMB=SCB/(r-1)$	
Entre tratamientos	SCTr	$r-1$	$CMTr=SCB/(r-1)$	$CMTr/CME$
Error	SCE	$(r-1)(r-2)$	$CME=SCE/(r-1)(r-2)$	
Total	SCT	r^2-1		

MODELOS ANCOVA DE LA COVARIANZA ANCOVA

Si ampliamos el análisis de la varianza suponiendo que influyen en la variable respuesta (variable independiente), además de los factores, una o varias variables cuantitativas, se aplicará un análisis de la covarianza ANCOVA para explicar correctamente dicha variable respuesta. Estas variables cuantitativas se denominan **covariantes** o **variables concomitantes**.

De una manera muy general puede considerarse que el análisis de la covarianza reúne las técnicas del análisis de la varianza y del análisis de la regresión. La diferencia entre análisis de la varianza y análisis de la regresión radica en la forma de tratar las variables independientes (factores). En el análisis de la regresión todos los factores son cuantitativos y se tratan cuantitativamente. En el análisis de la varianza los factores suelen ser cualitativos, pero si alguno es cuantitativo, se trata cualitativamente. En el análisis de la covarianza, por ser una mezcla de ambos análisis, unos factores se tratan cualitativamente (los factores del análisis de la varianza) y otros cuantitativamente (los covariantes).

Modelo con un factor y un covariante

El modelo de análisis de la covarianza más simple que se puede considerar es el que tiene un factor y un covariante, y será de la forma:

$$Y_{ij} = u + A_i + \beta X_{ij} + E_{ij} \quad i = 1, \dots, t \quad j = 1, \dots, n_i$$

donde A_i es el factor fijo y X_{ij} es el covariante. Notemos que X_{ij} no es una variable aleatoria. El error E_{ij} sí es una variable aleatoria, con las hipótesis de normalidad, homocedasticidad, independencia y esperanza matemática nula.

Las variables E_{ij} son normales $N(0, \sigma^2)$ e independientes, y por ser el factor A fijo, sus distintos niveles verificarán la condición $\sum_{i=1}^t n_i A_i = 0$.

Si $n_i = n$ para todo $i = 1, \dots, t$, el modelo es equilibrado.

Modelo con dos factores y un covariante

Para un modelo de dos factores y un covariante tendremos la expresión:

$$Y_{ij} = u + A_i + B_j + \beta X_{ij} + E_{ij} \quad i = 1, \dots, t, \quad j = 1, \dots, n_i$$

donde A_i y B_j son los factores fijos y X_{ij} es el covariante. Obsérvese que X_{ij} no es una variable aleatoria. El error E_{ij} sí es una variable aleatoria, con las hipótesis de normalidad, homocedasticidad, independencia y esperanza matemática nula.

Las variables E_{ij} son normales $N(0,\sigma)$ e independientes, y por ser los factores A y B fijos, sus distintos niveles verificarán la condición:

$$\sum_{i=1}^t A_i = \sum_{j=1}^n B_j = 0$$

Si consideramos $n_i = n$ para todo $i = 1, \dots, t$, el modelo es equilibrado.

Modelo con dos factores y dos covariantes

Para un modelo de dos factores y dos covariantes tendremos la expresión:

$$Y_{ij} = u + A_i + B_j + \gamma X_{ij} + \delta W_{ij} + E_{ij} \quad i = 1, \dots, t, \quad j = 1, \dots, n_i$$

donde A_i y B_j son los factores fijos y X_{ij} y W_{ij} son los covariantes. Observemos que X_{ij} y W_{ij} no son variables aleatorias. El error E_{ij} sí es una variable aleatoria, con las hipótesis de normalidad, homocedasticidad, independencia y esperanza matemática nula.

Las variables E_{ij} son normales $N(0,\sigma)$ e independientes, y por ser los factores A y B fijos, sus distintos niveles verificarán la condición:

$$\sum_{i=1}^t A_i = \sum_{j=1}^n B_j = 0 .$$

Si consideramos $n_i = n$ para todo $i = 1, \dots, t$, el modelo es equilibrado.

Modelo MANOVA (Análisis de la varianza múltiple)

El análisis de la varianza múltiple MANOVA es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medias de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza múltiple MANOVA es la siguiente:

$$G(y_1, y_2, \dots, y_m) = F(x_1, x_2, \dots, x_n)$$

Las variables dependientes son métricas y las variables independientes son no métricas.

Modelo MANCOVA (Análisis de la covarianza múltiple)

El análisis de la covarianza múltiple es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) mezcla de variables métricas y no métricas.

La expresión funcional del modelo del análisis de la covarianza múltiple MANCOVA es la siguiente:

$$G(y_1, y_2, \dots, y_m) = F(x_1, x_2, \dots, x_n)$$

Las variables dependientes son métricas y las variables independientes son una parte métricas y otra parte no métricas.

En el análisis de la covarianza, tanto simple como múltiple, las variables métricas independientes (*covariables*) tienen como objetivo eliminar determinados efectos que puedan sesgar los resultados incrementando la varianza dentro de los grupos. En el análisis de la covarianza se suele comenzar eliminando, mediante una regresión lineal, la variación experimentada por las variables dependientes producida por la covariable o covariables de efectos indeseados, para continuar con un análisis ANOVA o MANOVA sobre las variables dependientes ajustadas (residuos de la regresión anterior).

MODELO LINEAL GENERAL (GLM)

El modelo de regresión Múltiple Lineal General (GLM) es el modelo más general posible de regresión lineal, incluyendo el modelo de regresión lineal múltiple con variables cuantitativas y los modelos de regresión múltiple con variables cualitativas y cuantitativas a la vez, por lo que incluirá todos los modelos del análisis de la varianza y de la covarianza.

ANÁLISIS DE LA VARIANZA Y LA COVARIANZA CON SPSS

PROCEDIMIENTO ANOVA DE UN FACTOR

El procedimiento ANOVA de un factor genera un análisis de la varianza de un factor para una variable dependiente cuantitativa respecto a una única variable de factor (la variable independiente). El análisis de la varianza se utiliza para contrastar la hipótesis de que varias medias son iguales. Esta técnica es una extensión de la prueba *t* para dos muestras. Además de determinar que existen diferencias entre las medias, es posible que desee saber qué medias difieren. Existen dos tipos de contrastes para comparar medias: los contrastes *a priori* y las pruebas *post hoc*. Los contrastes *a priori* se plantean antes de ejecutar el experimento y las pruebas *post hoc* se realizan después de haber llevado a cabo el experimento. También puede contrastar las tendencias existentes a través de las categorías. En cuanto a estadísticos, para cada grupo se obtiene número de casos, media, desviación típica, error típico de la media, mínimo, máximo, intervalo de confianza al 95% para la media, prueba de Levene sobre la homogeneidad de varianzas, tabla de análisis de varianza para cada variable dependiente, contrastes *a priori* especificados por el usuario y las pruebas de rango y de comparaciones múltiples *post hoc*: Bonferroni, Sidak, diferencia honestamente significativa de Tukey, GT2 de Hochberg, Gabriel, Dunnett, prueba F de Ryan-Einot-Gabriel-Welsch (R-E-G-W F), prueba de rango de Ryan-Einot-Gabriel-Welsch (R-E-G-W Q), T2 de Tamhane, T3 de Dunnett, Games-Howell, C de Dunnett, prueba de rango múltiple de Duncan, Student-Newman-Keuls (S-N-K), Tukey b, Waller-Duncan, Scheffé y diferencia menos significativa.

Para obtener un análisis de varianza de un factor, elija en los menús *Analizar* → *Comparar medias* → *ANOVA de un factor* (Figura 18-1), seleccione una o más variables dependientes y seleccione una sola variable de factor independiente (Figura 18-2). En el fichero EMPLEADOS analizaremos el salario actual (*salario*) según el factor titulación mayor obtenida (*educ*).

El botón *Contrastes* (Figura 18-3) permite dividir las sumas de cuadrados intergrupos en componentes de tendencia o especificar contrastes *a priori*. En *Polinómico* se puede contrastar la existencia de tendencia en la variable dependiente a través de los niveles ordenados de la variable de factor. Por ejemplo, podría contrastar si existe una tendencia lineal (creciente o decreciente) en el salario, a través de los niveles ordenados de la titulación mayor obtenida. En *Orden* se puede elegir un orden polinómico 1º, 2º, 3º, 4º o 5º. En *Coeficientes* se pueden elegir contrastes *a priori* especificados por el usuario que serán contrastados mediante el estadístico *t*. Introduzca un coeficiente para cada grupo (categoría) de la variable factor y pulse en *Añadir después de cada entrada*. Cada nuevo valor se añade al final de la lista de coeficientes. Para especificar conjuntos de contrastes adicionales, pulse en *Siguiente*. Utilice *Siguiente* y *Previo* para desplazarse entre los conjuntos de contrastes.

El orden de los coeficientes es importante porque se corresponde con el orden ascendente de los valores de las categorías de la variable de factor. El primer coeficiente en la lista se corresponde con el menor de los valores de grupo en la variable de factor y el último coeficiente se corresponde con el valor más alto. Por ejemplo, si existen 10 categorías en la variable factor, los coeficientes -1, 0, 0, 0, 0, 5, 0, 5, 0, 0 y 0 contrastan el primer grupo con los grupos quinto y sexto. Para la mayoría de las aplicaciones, la suma de los coeficientes debería ser 0. Los conjuntos que no sumen 0 también se pueden utilizar, pero aparecerá un mensaje de advertencia.

Una vez que se ha determinado que existen diferencias entre las medias, las pruebas de rango *post hoc* (botón *Post hoc* de la Figura 18-2) y las comparaciones múltiples por parejas permiten determinar qué medias difieren. Las pruebas de rango *post hoc* (Figura 18-4) identifican subconjuntos homogéneos de medias que no se diferencian entre sí. Las comparaciones múltiples por parejas contrastan la diferencia entre cada pareja de medias y dan lugar a una matriz donde los asteriscos indican las medias de grupo significativamente diferentes a un nivel alfa de 0,05. La prueba de la diferencia honestamente significativa de Tukey, la GT2 de Hochberg, la prueba de Gabriel y la prueba de Scheffé son pruebas de comparaciones múltiples y pruebas de rango. Otras pruebas de rango disponibles son Tukey b, S-N-K (Student-Newman-Keuls), Duncan, R-E-G-W F (prueba F de Ryan-Einot-Gabriel-Welsch), R-E-G-W Q (prueba de rango de Ryan-Einot-Gabriel-Welsch) y Waller-Duncan. Las pruebas de comparaciones múltiples disponibles son Bonferroni, diferencia honestamente significativa de Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé, y DMS (diferencia menos significativa). Las pruebas de comparaciones múltiples que no suponen varianzas iguales son T2 de Tamhane, T3 de Dunnett, Games-Howell y C de Dunnett. Posiblemente le resulte más fácil interpretar el resultado de los contrastes *post hoc* si desactiva *Ocultar filas y columnas vacías* en el cuadro de diálogo *Propiedades de tabla* (en una tabla pivot activada, seleccione *Propiedades de tabla* en el menú *Formato*). El botón *Opciones* permite seleccionar *Estadísticos* y *Gráficos*. Al pulsar *Aceptar* en la Figura 18-2 se obtiene la salida (Figuras 10-5 a 10-7).

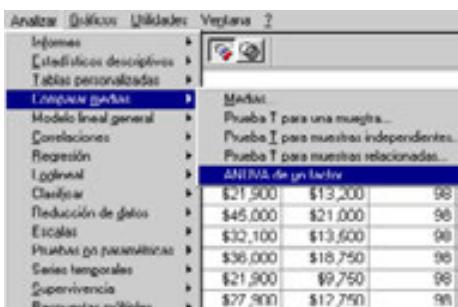


Figura 18-1

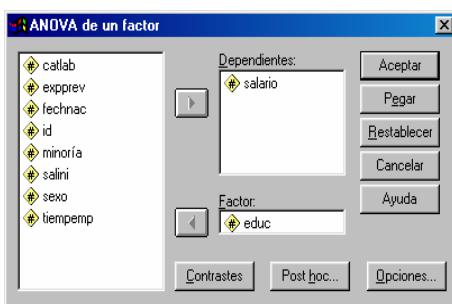


Figura 18-2

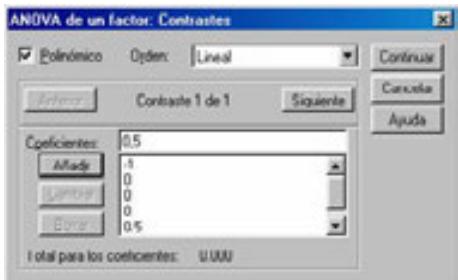


Figura 18-3

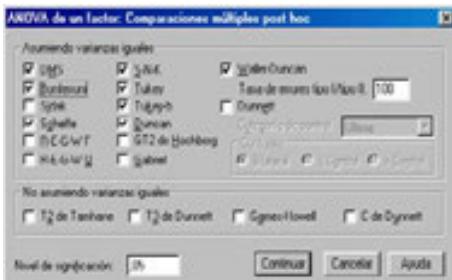


Figura 18-4

<pre> COMINCAT ANALISIS BY educ /POLYNOMIAL= 1 /CONTRAST= -1 0 0 0 0.5 0.5 0 0 0 /PRINTING ANALYSIS /METHODS = BHK TUKEY BTUKEY DUNCAN SCHEFFE LOD BONFERRONI NELLER (100) ALPHA (.05). </pre>																																										
ANOVA de un factor																																										
Salario actual																																										
<table border="1"> <thead> <tr> <th></th> <th></th> <th>Suma de cuadrados</th> <th>gl</th> <th>Media cuadrática</th> <th>F</th> <th>Sig.</th> </tr> </thead> <tbody> <tr> <td>Inter-grupos</td> <td>Combinado-s2</td> <td>0,00E+00</td> <td>9</td> <td>9,00E+09</td> <td>92,779</td> <td>,000</td> </tr> <tr> <td></td> <td>Término lineal</td> <td>6,014E+10</td> <td>1</td> <td>6,02E+10</td> <td>566,909</td> <td>,000</td> </tr> <tr> <td></td> <td>Ponderado</td> <td>2,040E+10</td> <td>0</td> <td>3,56E+09</td> <td>33,526</td> <td>,000</td> </tr> <tr> <td>Intra-grupos</td> <td>Desviación</td> <td>-4,926E+10</td> <td>464</td> <td>1,06E+08</td> <td></td> <td></td> </tr> <tr> <td>Total</td> <td></td> <td>1,273E+11</td> <td>473</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>			Suma de cuadrados	gl	Media cuadrática	F	Sig.	Inter-grupos	Combinado-s2	0,00E+00	9	9,00E+09	92,779	,000		Término lineal	6,014E+10	1	6,02E+10	566,909	,000		Ponderado	2,040E+10	0	3,56E+09	33,526	,000	Intra-grupos	Desviación	-4,926E+10	464	1,06E+08			Total		1,273E+11	473			
		Suma de cuadrados	gl	Media cuadrática	F	Sig.																																				
Inter-grupos	Combinado-s2	0,00E+00	9	9,00E+09	92,779	,000																																				
	Término lineal	6,014E+10	1	6,02E+10	566,909	,000																																				
	Ponderado	2,040E+10	0	3,56E+09	33,526	,000																																				
Intra-grupos	Desviación	-4,926E+10	464	1,06E+08																																						
Total		1,273E+11	473																																							

Figura 18-5

Coeficientes de los contrastes										
Contraste	Nivel educativo									
	0	12	14	15	16	17	10	19	20	21
1	-1	0	0	0	,5	,5	0	0	0	0

Pruebas para los contrastes						
	Contraste	Valor del contraste	Error típico	t	gl	Sig. (doblez)
Salario actual: Admitiendo igualdad de varianzas	1	\$29,477,55	\$2,205,91	13,363	464	,000
No asumiendo	1	\$29,477,55	\$2,205,91	13,363	464	,000

Figura 18-6

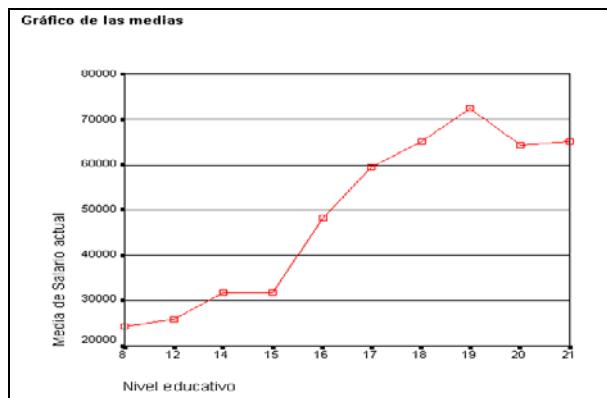


Figura 18-7

PROCEDIMIENTO MLG UNIVARIANTE

El procedimiento MLG Univariante proporciona un análisis de regresión y un análisis de varianza para una variable dependiente mediante uno o más factores o variables. Las variables de factor dividen la población en grupos. Con el procedimiento MLG (modelo lineal general) se pueden contrastar hipótesis nulas sobre los efectos de otras variables en las medias de varias agrupaciones de una única variable dependiente. Se pueden investigar las interacciones entre los factores así como los efectos de los factores individuales, algunos de los cuales pueden ser aleatorios. Además, se pueden incluir los efectos de las covariables y las interacciones de covariables con los factores. Para el análisis de regresión, las variables independientes (predictoras) se especifican como covariables.

Se pueden contrastar tanto los modelos equilibrados como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. Además de contrastar hipótesis, MLG Univariante genera estimaciones de los parámetros. También se encuentran disponibles los contrastes *a priori* de uso más habitual para contrastar las hipótesis. Además, si una prueba F global ha mostrado cierta significación, pueden emplearse las pruebas *post hoc* para evaluar las diferencias entre las medias específicas. Las medias marginales estimadas ofrecen estimaciones de valores de las medias pronosticadas para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones. En su archivo de datos puede guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos. Ponderación MCP permite especificar una variable usada para aplicar a las observaciones una ponderación diferente en un análisis de mínimos cuadrados ponderados (MCP), por ejemplo para compensar la distinta precisión de las medidas.

En cuanto a estadísticos, se obtienen las pruebas de rango *post hoc* y las comparaciones múltiples: diferencia menos significativa (DMS), Bonferroni, Sidak, Scheffé, múltiples F de Ryan-Einot-Gabriel-Welsch (R-E-G-WF), rango múltiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls (S-N-K), diferencia honestamente significativa de Tukey, b de Tukey, Duncan, GT2 de Hochberg, Gabriel, pruebas *t* de Waller Duncan, Dunnett (unilateral y bilateral), T2 de Tamhane, T3 de Dunnett, Games-Howell y C de Dunnett. Estadísticos descriptivos: medias observadas, desviaciones típicas y frecuencias de todas las variables dependientes en todas las casillas. Prueba de Levene para la homogeneidad de varianzas. En cuanto a gráficos se obtienen diagramas de dispersión por nivel, gráficos de residuos y gráficos de perfil (interacción).

Para obtener un análisis MLG Univariante, elija en los menús *Analizar → Modelo lineal general → Univariante* (Figura 18-8), seleccione una variable dependiente, seleccione variables para *Factores fijos*, *Factores aleatorios* y *Covariables*, en función de los datos (Figura 18-9). Para especificar una variable de ponderación, utilice *Ponderación MCP*. El botón *Especificar modelo* (Figura 18-10) permite definir un modelo factorial completo que contiene todos los efectos principales del factor, todos los efectos principales de las covariables y todas las interacciones factor por factor. No contiene interacciones de covariable. Seleccione *Personalizado* para especificar sólo un subconjunto de interacciones o para especificar interacciones factor por covariable. Indique todos los términos que deseé incluir en el modelo. Como ejemplo, en el fichero EMPLEADOS usaremos como variable dependiente *salario*, como factor fijo *sexo*, como factor aleatorio *catlab* y como covariables *tiempemp*, *exprev* y *salini*.

El botón *Contrastes* (Figura 18-11) permite definir los contrastes de las diferencias entre los niveles de un factor. Puede especificar un contraste para cada factor en el modelo (en un modelo de medidas repetidas, para cada factor inter-sujetos). Los contrastes representan las combinaciones lineales de los parámetros.

El botón *Gráficos* (Figura 18-12) permite definir los gráficos de perfil (gráficos de interacción) que sirven para comparar las medias marginales en el modelo. Un gráfico de perfil es un gráfico de líneas en el que cada punto indica la media marginal estimada de una variable dependiente (corregida respecto a las covariables) en un nivel de un factor. Los niveles de un segundo factor se pueden utilizar para generar líneas diferentes. Cada nivel en un tercer factor se puede utilizar para crear un gráfico diferente. Todos los factores fijos y aleatorios, si existen, están disponibles para los gráficos. Para los análisis multivariados, los gráficos de perfil se crean para cada variable dependiente. En un análisis de medidas repetidas, es posible utilizar tanto los factores inter-sujetos como los intra-sujetos en los gráficos de perfil. Las opciones *MLG - Multivariante* y *MLG - Medidas repetidas* sólo estarán disponibles si tiene instalada la opción *Modelos avanzados*. Un gráfico de perfil de un factor muestra si las medias marginales estimadas aumentan o disminuyen a través de los niveles. Para dos o más factores, las líneas paralelas indican que no existe interacción entre los factores, lo que significa que puede investigar los niveles de un único factor. Las líneas no paralelas indican una interacción.

El botón *Post Hoc* ya fue explicado en el procedimiento anterior. El botón *Opciones* (Figura 18-13) permite seleccionar estadísticos adicionales. El botón *Guardar* permite guardar los valores pronosticados por el modelo, los residuos y las medidas relacionadas como variables nuevas en el editor de datos. Muchas de estas variables se pueden utilizar para examinar supuestos sobre los datos. Si desea almacenar los valores para utilizarlos en otra sesión de SPSS, guárdelos en el archivo de datos actual. Al pulsar *Aceptar* en la Figura 18-9 se obtiene la salida (Figuras 10-14 a 10-31).

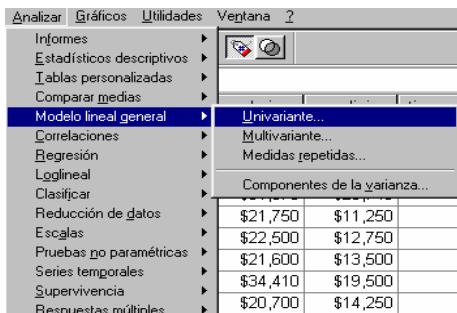


Figura 18-8

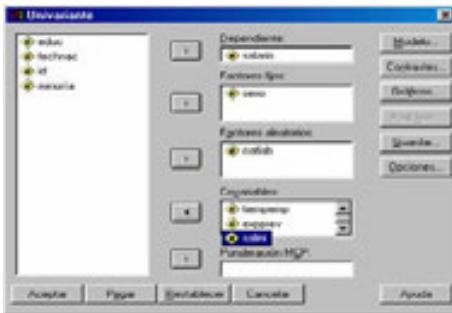


Figura 18-9

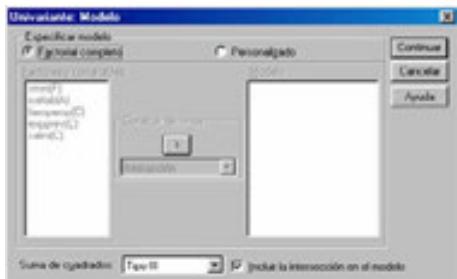


Figura 18-10



Figura 18-11



Figura 18-12

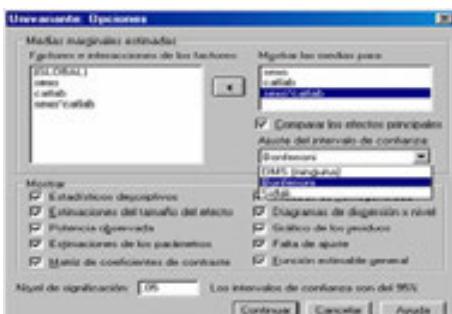


Figura 18-13

```

UNIANOVA
  salario BY sexo catlab WITH tiempemp expprev salini
  /RANDOM = catlab
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /PLOT = PROFILE( catlab sexo )
  /EMMEANS = TABLES(sexo) WITH(tiempemp=MEAN expprev=MEAN salini=MEAN)
  COMPARE ADJ(BONFERRONI)
  /EMMEANS = TABLES(catlab) WITH(tiempemp=MEAN expprev=MEAN salini=MEAN)
  COMPARE ADJ(BONFERRONI)
  /EMMEANS = TABLES(sexo*catlab) WITH(tiempemp=MEAN expprev=MEAN salini=MEAN)
  /PRINT = DESCRIPTIVE ETASQ OPOWER PARAMETER TEST(LMATRIX) HOMOGENEITY LOF GEF
  /PLOT = SPREADLEVEL RESIDUALS
  /CRITERIA = ALPHA(.05)
  /DESIGN = tiempemp expprev salini sexo catlab sexo*catlab .

```

Factores inter-sujetos

		Etiqueta del valor	N
Sexo	1	Hombre	258
	2	Mujer	216
Categoría laboral	1	Administrativo	363
	2	Seguridad	27
	3	Directivo	84

Figura 18-14

Estadísticos descriptivos				
Variable dependiente: Salario actual				
Sexo	Categoría laboral	Media	Desv. tipo	N
Hombre	Administrativo	\$7,997,98	157	
	Seguridad	\$2,114,62	27	
	Directivo	-----	74	
	Total	-----	258	
Mujer	Administrativo	\$5,812,84	206	
	Directivo	\$8,501,25	10	
	Total	\$7,558,02	216	
	Total	\$7,567,99	363	
Total	Administrativo	\$2,114,62	27	
	Seguridad	-----	84	
	Directivo	-----	474	
	Total	-----	-----	

Contraste de Levene sobre la igualdad de las varianzas error^a

Variable dependiente: Salario actual			
F	gl1	gl2	Significación
32,767	4	469	.000

Contraste la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diccionario: Intercept+TIEMPEMP+EXPprev+SALINI+SEXO+CATLAB+SEXO * CATLAB

Figura 18-15

Media cuadrática esperada ^{a,b}				
Fuente	Componente de la varianza			
	Var(CATLAB)	Var(SEXO * CATLAB)	Var(Error)	Término cuadrático
Intersección	1,776	1,098	1,000	Intercept, SEXO
TIEMPEMP	,000	,000	1,000	TIEMPEM
EXPprev	,000	,000	1,000	EXPPRE
SALINI	,000	,000	1,000	SALINI
SEXO	,000	26,792	1,000	SEXO
CATLAB	36,232	24,434	1,000	
SEXO * CATLAB	,000	30,072	1,000	
Error	,000	,000	1,000	

a. Para cada fuente, la media cuadrática esperada es igual a la suma de los coeficientes de las casillas por las componentes de la varianza, más un término cuadrático que incluye los efectos de la casilla Término cuadrático.
b. Las medias cuadráticas esperadas se basan en la suma de cuadrados tipo III.

Figura 18-16

Pruebas de los efectos inter-sujetos								
Variable dependiente: Salario actual								
Fuente	Suma de cuadrados tipo III	gl	Media cuadráti ca	F	Significa ción	Eta cuadrad o	Parámet ro de no centralid ad	Potencia observada ^a
Intersección	Hipótesis	7,1E+07	1	7,1E+07	,558	,490	,104	,558 ,094
	Error	6,1E+08	4,813	1,3E+08 ^b				
TIEMPEMP	Hipótesis	1,1E+09	1	1,1E+09	23,882	,000	,049	23,882 ,998
	Error	2,2E+10	466	4,7E+07 ^c				
EXPprev	Hipótesis	2,4E+09	1	2,4E+09	50,959	,000	,099	50,959 1,000
	Error	2,2E+10	466	4,7E+07 ^c				
SALINI	Hipótesis	1,8E+10	1	1,8E+10	379,935	,000	,449	379,935 1,000
	Error	2,2E+10	466	4,7E+07 ^c				
SEXO	Hipótesis	1,7E+08	1	1,7E+08	30,600	,000	,138	30,600 1,000
	Error	1,0E+09	190,610	5485668 ^d				
CATLAB	Hipótesis	3,3E+09	2	1,7E+09	182,048	,000	,523	364,096 1,000
	Error	3,1E+09	332,615	9173811 ^e				
SEXO * CATLAB	Hipótesis	355907	1	355907	,008	,931	,000	,008 ,051
	Error	2,2E+10	466	4,7E+07 ^c				

a. Calculado con alfa = ,05
b. 4,903E-02 MS(CATLAB) - 3,322E-03 MS(SEXO * CATLAB) + ,954 MS(Error)
c. MS(Error)
d. ,891 MS(SEXO * CATLAB) + ,109 MS(Error)
e. ,813 MS(SEXO * CATLAB) + ,187 MS(Error)

Figura 18-17

Estimaciones de los parámetros									
Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%		Eta cuadrado	Parámetro de no centralidad	Potencia observada ^a
					Límite inferior	Límite superior			
Intersección	8401,72	3639,34	2,309	,021	1250,17	15553,3	,011	2,309	,635
TIEMPEMP	154,821	31,681	4,887	,000	92,566	217,075	,049	4,887	,998
EXPPREV	-24,969	3,498	-7,139	,000	-31,843	-18,096	,099	7,139	1,000
SALINI	1,410	,072	19,492	,000	1,267	1,552	,449	19,492	1,000
[SEXO=1]	2611,86	2470,95	1,057	,291	-2243,7	7467,45	,002	1,057	,184
[SEXO=2]	^b								
[CATLAB=1]	-11895	2297,19	-5,178	,000	-16409	-7380,5	,054	5,178	,999
[CATLAB=2]	-6509,8	2166,27	-3,005	,003	-10767	-2253,0	,019	3,005	,851
[CATLAB=3]	^b								
[SEXO=1] * [CATLAB=1]	-217,580	2510,81	-,087	,931	-5151,5	4716,34	,000	,087	,051
[SEXO=1] * [CATLAB=2]	^b								
[SEXO=1] * [CATLAB=3]	^b								
[SEXO=2] * [CATLAB=1]	^b								
[SEXO=2] * [CATLAB=3]	^b								

a. Calculado con alfa = ,05
b. Al parámetro se le ha asignado el valor cero porque es redundante.

Figura 18-18

Parámetro	Función estimable general ^a								
	L1	L2	L3	L4	L5	L7	L8	L10	
Intersección	1,000	,000	,000	,000	,000	,000	,000	,000	,000
TIEMPEMP	,000	1,000	,000	,000	,000	,000	,000	,000	,000
EXPPREV	,000	,000	1,000	,000	,000	,000	,000	,000	,000
SALINI	,000	,000	,000	1,000	,000	,000	,000	,000	,000
[SEXO=1]	,000	,000	,000	,000	1,000	,000	,000	,000	,000
[SEXO=2]	1,000	,000	,000	,000	,000	-1,000	,000	,000	,000
[CATLAB=1]	,000	,000	,000	,000	,000	,000	1,000	,000	,000
[CATLAB=2]	,000	,000	,000	,000	,000	,000	,000	1,000	,000
[CATLAB=3]	1,000	,000	,000	,000	,000	,000	-1,000	-1,000	,000
[SEXO=1] * [CATLAB=1]	,000	,000	,000	,000	,000	,000	,000	,000	1,000
[SEXO=1] * [CATLAB=2]	,000	,000	,000	,000	,000	,000	,000	1,000	,000
[SEXO=1] * [CATLAB=3]	,000	,000	,000	,000	,000	1,000	,000	-1,000	,000
[SEXO=2] * [CATLAB=1]	,000	,000	,000	,000	,000	,000	1,000	,000	-1,000
[SEXO=2] * [CATLAB=3]	1,000	,000	,000	,000	,000	-1,000	-1,000	,000	1,000

a. Diseño: Intercept+TIEMPEMP+EXPPREV+SALINI+SEXO+CATLAB+SEXO * CATLAB

Figura 18-19

Pruebas de falta de ajuste								
Variable dependiente: Salario actual								
Fuente	Suma de cuadrados	gl	Media cuadrática	F	Significación	Eta cuadrado	Parámetro de no centralidad	Potencia observada ^a
Falta de ajuste	2,207E+10	461	47883840	21,127	,001	,999	9739,444	1,003
Error puro	11332500	5	2266500,0					

a. Calculado con alfa = ,05

Figura 18-20

Medias marginales estimadas**1. Sexo****Coeficientes de contraste (matriz L*)**

Parámetro	Sexo	
	Hombre	Mujer
Intersección	1	1
TIEMPEMP	81,110	81,110
EXPprev	95,861	95,861
SALINI	17016,086	17016,086
[SEXO=1]	1	0
[SEXO=2]	0	1
[CATLAB=1]	,333	,500
[CATLAB=2]	,333	0
[CATLAB=3]	,333	,500
[SEXO=1] * [CATLAB=1]	,333	0
[SEXO=1] * [CATLAB=2]	,333	0
[SEXO=1] * [CATLAB=3]	,333	0
[SEXO=2] * [CATLAB=1]	0	,500
[SEXO=2] * [CATLAB=3]	0	,500

Figura 18-21

2. Categoría laboral**Coeficientes de contraste (matriz L*)**

Parámetro	Categoría laboral		
	Administrativo	Seguridad	Directivo
Intersección	1	1	1
TIEMPEMP	81,110	81,110	81,110
EXPprev	95,861	95,861	95,861
SALINI	17016,086	17016,086	17016,086
[SEXO=1]	,500	1	,500
[SEXO=2]	,500	0	,500
[CATLAB=1]	1	0	0
[CATLAB=3]	0	0	1
[SEXO=1] * [CATLAB=1]	,500	0	0
[SEXO=1] * [CATLAB=2]	0	1	0
[SEXO=1] * [CATLAB=3]	0	0	,500
[SEXO=2] * [CATLAB=1]	,500	0	0
[SEXO=2] * [CATLAB=3]	0	0	,500

Figura 18-22

Estimaciones

Variable dependiente: Salario actual

Sexo	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
Hombre	38954,489 ^a	637,091	37702,562	40206,416
Mujer	36602,662 ^a	1122,385	34397,098	38808,225

a. Evaluado respecto a cómo aparecen las covariables en el modelo: Meses desde el contrato = 81,11, Experiencia previa (meses) = 95,86, Salario inicial = \$17,016,09.

b. Basada en la media marginal poblacional modificada.

Figura 18-23

Estimaciones

Variable dependiente: Salario actual

Categoría laboral	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
Administrativo	31852,495 ^a	412,844	31041,229	32663,761
Seguridad	38651,995 ^a	1518,607	35667,830	41636,161
Directivo	43855,903 ^a	1334,814	41233,296	46478,510

a. Evaluado respecto a cómo aparecen las covariables en el modelo: Meses desde el contrato = 81,11, Experiencia previa (meses) = 95,86, Salario inicial = \$17,016,09.

b. Basada en la media marginal poblacional modificada.

Figura 18-24

Comparaciones por pares

Variable dependiente: Salario actual

(I) Sexo	(J) Sexo	Diferencia entre medias (I-J)	Error típ.	Significación ^a	Intervalo de confianza al 95% para diferencia ^a	
					Límite inferior	Límite superior
Hombre	Mujer	2351,827 ^b	1314,191	,074	-230,648	4934,302
Mujer	Hombre	-2351,827 ^c	1314,191	,074	-4934,302	230,648

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

b. Una estimación de la media marginal poblacional modificada (J).

c. Una estimación de la media marginal poblacional modificada (I).

Figura 18-25

Comparaciones por pares

Variable dependiente: Salario actual

(I) Categoría laboral	(J) Categoría laboral	Diferencia entre medias (I-J)	Error típ.	Significación ^a	Intervalo de confianza al 95% para diferencia ^a	
					Límite inferior	Límite superior
Administrativo	Seguridad	-6799,501 ^a	1558,378	,000	-10543,744	-3055,257
	Directivo	-12003,408 ^a	1481,015	,000	-15561,776	-8445,041
Seguridad	Administrativo	6799,501 ^a	1558,378	,000	3055,257	10543,744
	Directivo	-5203,908 ^a	2132,288	,045	-10327,059	-80,757
Directivo	Administrativo	12003,408 ^a	1481,015	,000	8445,041	15561,776
	Seguridad	5203,908 ^a	2132,288	,045	80,757	10327,059

Basadas en las medias marginales estimadas.

* La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

b. Una estimación de la media marginal poblacional modificada (J).

c. Una estimación de la media marginal poblacional modificada (I).

Figura 18-26

Contrastes entre los factores								
Variable dependiente: Salario actual								
	Suma de cuadrados	gl	Media cuadrática	F	Significación	Eta cuadrado	Parámetro de no centralidad	Potencia observada ^a
Contraste	15178,2110	1	1,52E+08	3,203	,074	,007	3,203	,431
Error	2,209E+10	466	47394,384					

Cada prueba F contrasta el efecto simple de Sexo en cada combinación de niveles del resto de los efectos mostrados. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas.

a. Calculado con alfa = ,05

Figura 18-27

Contrastes entre los factores								
Variable dependiente: Salario actual								
	Suma de cuadrados	gl	Media cuadrática	F	Significación	Eta cuadrado	Parámetro de no centralidad	Potencia observada ^a
Contraste	3,962E+09	2	1,98E+09	41,797	,000	,152	83,595	1,000
Error	2,209E+10	466	47394,384					

Cada prueba F contrasta el efecto simple de Categoría laboral en cada combinación de niveles del resto de los efectos mostrados. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas.

a. Calculado con alfa = ,05

Figura 18-28

3. Sexo * Categoría laboral

Coeficientes de contraste (matriz L')

Parámetro	Sexo					
	Hombre			Mujer		
	Categoría laboral			Categoría laboral		
Administrativo	Seguridad	Directivo	Administrativo	Directivo		
Intersección	1	1	1	1	1	1
TIEMPEMP	81,110	81,110	81,110	81,110	81,110	81,110
EXPRESS	95,861	95,861	95,861	95,861	95,861	95,861
SALINI	17016,086	17016,086	17016,086	17016,086	17016,086	17016,086
[SEXO=1]	1	1	0	0	0	0
[SEXO=2]	0	0	0	1	1	1
[CATLAB=1]	1	0	0	1	0	0
[CATLAB=2]	0	1	0	0	0	0
[CATLAB=3]	0	0	1	0	1	1
[SEXO=1] * [CATLAB=1]	1	0	0	0	0	0
[SEXO=1] * [CATLAB=2]	0	1	0	0	0	0
[SEXO=1] * [CATLAB=3]	0	0	1	0	0	0
[SEXO=2] * [CATLAB=1]	0	0	0	1	0	0
[SEXO=2] * [CATLAB=3]	0	0	0	0	1	1

No se muestran las combinaciones de los niveles sin observaciones.

Figura 18-29

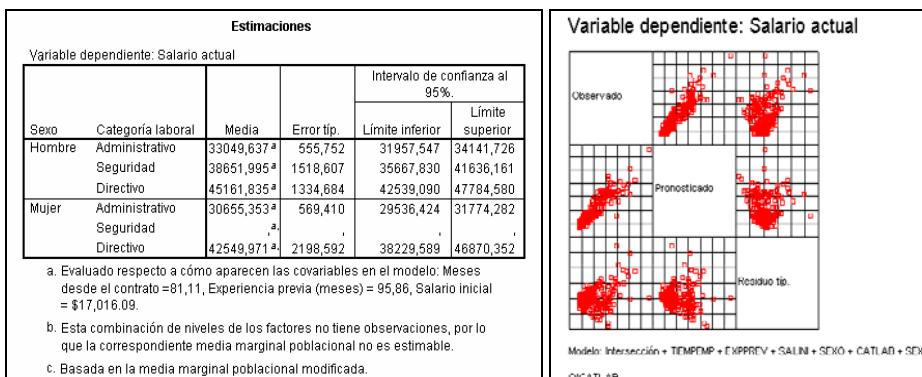


Figura 18-31

PROCEDIMIENTO MLG MULTIVARIANTE

El procedimiento MLG Multivariante proporciona un análisis de regresión y un análisis de varianza para variables dependientes múltiples por una o más covariables o variables de factor. Las variables de factor dividen la población en grupos. Utilizando este procedimiento del modelo lineal general, es posible contrastar hipótesis nulas sobre los efectos de las variables de factor sobre las medias de varias agrupaciones de una distribución conjunta de variables dependientes. Asimismo, puede investigar las interacciones entre los factores y también los efectos individuales de los factores. Además, se pueden incluir los efectos de las covariables y las interacciones de covariables con los factores. Para el análisis de regresión, las variables independientes (predictoras) se especifican como covariables.

Se pueden contrastar tanto los modelos equilibrados como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. En un modelo multivariado, las sumas de cuadrados debidas a los efectos del modelo y las sumas de cuadrados error se encuentran en forma de matriz en lugar de en la forma escalar del análisis univariado. Estas matrices se denominan matrices SCPC (sumas de cuadrados y productos cruzados). Si se especifica más de una variable dependiente, se proporciona el análisis multivariado de varianzas usando la traza de Pillai, la lambda de Wilks, la traza de Hotelling y el criterio de mayor raíz de Roy con el estadístico F aproximado, así como el análisis univariado de varianza para cada variable dependiente. Además de contrastar hipótesis, MLG Multivariante genera estimaciones de los parámetros. También se encuentran disponibles los contrastes *a priori* de uso más habitual para contrastar las hipótesis. Además, si una prueba F global ha mostrado cierta significación, pueden emplearse las pruebas *post hoc* para evaluar las diferencias entre las medias específicas. Las medias marginales estimadas ofrecen estimaciones de valores de las medias pronosticados para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones. Las pruebas de comparaciones múltiples *post hoc* se realizan por separado para cada variable dependiente.

En su archivo de datos puede guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos. También se hallan disponibles una matriz SCPC residual, que es una matriz cuadrada de las sumas de cuadrados y los productos cruzados de los residuos; una matriz de covarianza residual, que es la matriz SCPC residual dividida por los grados de libertad de los residuos; y la matriz de correlaciones residual, que es la forma tipificada de la matriz de covarianza residual. Ponderación MCP permite especificar una variable usada para aplicar a las observaciones una ponderación diferencial en un análisis de mínimos cuadrados ponderados (MCP), por ejemplo para compensar la distinta precisión de las medidas.

En cuanto a estadísticos se obtienen las pruebas de rango *post hoc* y las comparaciones múltiples, diferencia menos significativa (DMS), Bonferroni, Sidak, Scheffé, múltiples F de Ryan-Einot-Gabriel-Welsch (R-E-G-W-F), rango múltiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls (S-N-K), diferencia honestamente significativa de Tukey, b de Tukey, Duncan, GT2 de Hochberg, Gabriel, pruebas t de Waller Duncan, Dunnett (unilateral y bilateral), T2 de Tamhane, T3 de Dunnett, Games-Howell y C de Dunnett, estadísticos descriptivos, medias observadas, desviaciones típicas y recuentos de todas las variables dependientes en todas las casillas; la prueba de Levene sobre la homogeneidad de la varianza; la prueba M de Box sobre la homogeneidad de las matrices de covarianza de las variables dependientes; y la prueba de esfericidad de Bartlett. En cuanto a gráficos se obtienen diagramas de dispersión por nivel, gráficos de residuos y gráficos de perfil (interacción).

Para obtener un análisis de varianza MLG Multivariante, elija en los menús *Analizar* → *Modelo lineal general* → *Multivariante* (Figura 18-32) y seleccione al menos dos variables dependientes. Si lo desea, puede especificar *Factores fijos*, *Covariables* y *Ponderación MCP* (Figura 18-33). Los botones de la Figura 18-33 funcionan como en el MLG Univariante. Usaremos el ejemplo del procedimiento anterior añadiendo *tiemprev* como variable dependiente. La salida se ve en las Figuras 10-34 a 10-36.

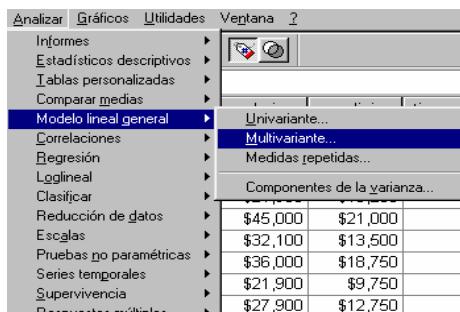


Figura 18-32



Figura 18-33

GLM		
	salario tiempesp	BY catlab sexo WITH expprev salini
/METHOD = SSTYPE(3)		
/INTERCEPT = INCLUDE		
/PLOT = PROFILE(catlab sexo)		
/CRITERIA = ALPHA(.05)		
/DESIGN = expprev salini catlab sexo catlab*sexo .		
Modelo lineal general		
Factores inter-sujetos		
Categoría	Etiqueta del valor	N
laboral	Administrativo	363
	Seguridadi	27
	Directivo	84
Sexo	Hombre	258
	Mujer	216

Figura 18-34

Contrastes multivariados ^c						
Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación
Intercept	Traza de Pillai	,767	765,354 ^a	2,000	466,000	,000
	Lambda de Wilks	,233	765,354 ^a	2,000	466,000	,000
	Traza de Hotelling	3,285	765,354 ^a	2,000	466,000	,000
	Raíz mayor de Roy	3,285	765,354 ^a	2,000	466,000	,000
EXPPREV	Traza de Pillai	,099	25,483 ^a	2,000	466,000	,000
	Lambda de Wilks	,901	25,483 ^a	2,000	466,000	,000
	Traza de Hotelling	,109	25,483 ^a	2,000	466,000	,000
	Raíz mayor de Roy	,109	25,483 ^a	2,000	466,000	,000
SALINI	Traza de Pillai	,453	192,578 ^a	2,000	466,000	,000
	Lambda de Wilks	,547	192,578 ^a	2,000	466,000	,000
	Traza de Hotelling	,827	192,578 ^a	2,000	466,000	,000
	Raíz mayor de Roy	,827	192,578 ^a	2,000	466,000	,000
CATLAB	Traza de Pillai	,133	16,568	4,000	934,000	,000
	Lambda de Wilks	,868	17,142 ^a	4,000	932,000	,000
	Traza de Hotelling	,152	17,715	4,000	930,000	,000
	Raíz mayor de Roy	,151	35,324 ^b	2,000	467,000	,000
SEXO	Traza de Pillai	,019	4,457 ^a	2,000	466,000	,012
	Lambda de Wilks	,981	4,457 ^a	2,000	466,000	,012
	Traza de Hotelling	,019	4,457 ^a	2,000	466,000	,012
	Raíz mayor de Roy	,019	4,457 ^a	2,000	466,000	,012
CATLAB * SEXO	Traza de Pillai	,005	1,133 ^a	2,000	466,000	,323
	Lambda de Wilks	,995	1,133 ^a	2,000	466,000	,323
	Traza de Hotelling	,005	1,133 ^a	2,000	466,000	,323
	Raíz mayor de Roy	,005	1,133 ^a	2,000	466,000	,323

Figura 18-35

Pruebas de los efectos inter-sujetos						
Fuente	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	Salario actual	1,147E+11 ^a	6	1,91E+10	384,509	,000
	Meses desde el contrato	657,368 ^b	6	109,561	1,084	,371
Intercept	Salario actual	6124104214	1	6,12E+09	123,180	,000
	Meses desde el contrato	154462,280	1	154462,280	1527,583	,000
EXPPREV	Salario actual	2403589829	1	2,40E+09	48,346	,000
	Meses desde el contrato	,585	1	,585	,006	,939
SALINI	Salario actual	1,741E+10	1	1,74E+10	350,253	,000
	Meses desde el contrato	291,444	1	291,444	2,882	,090
CATLAB	Salario actual	3316663989	2	1,66E+09	33,356	,000
	Meses desde el contrato	53,988	2	26,994	,267	,766
SEXO	Salario actual	276522092	1	2,77E+08	5,562	,019
	Meses desde el contrato	540,335	1	540,335	5,344	,021
CATLAB * SEXO	Salario actual	8647305,957	1	8647306,0	,174	,677
	Meses desde el contrato	228,950	1	228,950	2,264	,133
Error	Salario actual	2,322E+10	467	49716575		
	Meses desde el contrato	47220,927	467	101,115		
Total	Salario actual	6,995E+11	474			
	Meses desde el contrato	3166222,000	474			
Total corregida	Salario actual	1,379E+11	473			
	Meses desde el contrato	47878,295	473			

a. R cuadrado = ,832 (R cuadrado corregida = ,829)
b. R cuadrado = ,014 (R cuadrado corregida = ,001)

Figura 18-36

PROCEDIMIENTO MLG MEDIDAS REPETIDAS

El procedimiento MLG Medidas repetidas proporciona un análisis de varianza cuando se toma la misma medida varias veces a cada sujeto o caso. Si se especifican factores inter-sujetos, éstos dividen la población en grupos. Utilizando este procedimiento del modelo lineal general, puede contrastar hipótesis nulas sobre los efectos tanto de los factores inter-sujetos como de los factores intra-sujetos. Asimismo puede investigar las interacciones entre los factores y también los efectos individuales de los factores. También se pueden incluir los efectos de covariables constantes y de las interacciones de las covariables con los factores inter-sujetos.

En un diseño doblemente multivariado de medidas repetidas, las variables dependientes representan medidas de más de una variable para los diferentes niveles de los factores intra-sujetos. Por ejemplo, se pueden haber medido el pulso y la respiración de cada sujeto en tres momentos diferentes. El procedimiento MLG Medidas repetidas ofrece análisis univariados y multivariados para datos de medidas repetidas. Se pueden contrastar tanto los modelos equilibrados como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. En un modelo multivariado, las sumas de cuadrados debidas a los efectos del modelo y las sumas de cuadrados error se encuentran en forma de matriz en lugar de en la forma escalar del análisis univariado. Estas matrices se denominan matrices SCPC (sumas de cuadrados y productos cruzados). Además de contrastar las hipótesis, MLG Medidas repetidas genera estimaciones de los parámetros.

Se encuentran disponibles los contrastes *a priori* utilizados habitualmente para elaborar hipótesis que contrastan los factores inter-sujetos. Además, si una prueba F global ha mostrado cierta significación, pueden emplearse las pruebas *post hoc* para evaluar las diferencias entre las medias específicas. Las medias marginales estimadas ofrecen estimaciones de valores de las medias pronosticados para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones. En su archivo de datos puede guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos. También se hallan disponibles una matriz SCPC residual, que es una matriz cuadrada de las sumas de cuadrados y los productos cruzados de los residuos; una matriz de covarianza residual, que es la matriz SCPC residual dividida por los grados de libertad de los residuos; y la matriz de correlaciones residual, que es la forma tipificada de la matriz de covarianza residual. Ponderación MCP permite especificar una variable usada para aplicar a las observaciones una ponderación diferencial en un análisis de mínimos cuadrados ponderados (MCP), por ejemplo para compensar la distinta precisión de las medidas.

En cuanto a estadísticos se obtienen pruebas de rango *post hoc* y comparaciones múltiples (para los factores inter-sujetos): diferencia menos significativa (DMS), Bonferroni, Sidak, Scheffé, múltiples F de Ryan-Einot-Gabriel-Welsch (R-E-G-WF), rango múltiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls (S-N-K), diferencia honestamente significativa de Tukey, b de Tukey, Duncan, GT2 de Hochberg, Gabriel, pruebas t de Waller Duncan, Dunnett (unilateral y bilateral), T2 de Tamhane, T3 de Dunnett, Games-Howell y C de Dunnett. Se obtienen como estadísticos descriptivos: medias observadas, desviaciones típicas y recuentos de todas las variables dependientes en todas las casillas; la prueba de Levene sobre la homogeneidad de la varianza; la M de Box; y la prueba de esfericidad de Mauchly. En cuanto a gráficos se obtienen diagramas de dispersión por nivel, gráficos de residuos y gráficos de perfil (interacción).

Para obtener un análisis MLG de Medidas repetidas, elija en los menús *Analizar* → *Modelo lineal general* → *Medidas repetidas* (Figura 18-37), defina al menos un factor intra-sujeto y su número de niveles (Figura 18-38) y pulse en *Definir*. Seleccione en la lista una variable dependiente que corresponda a cada combinación de factores intra-sujetos (y, de forma opcional, medidas). Para cambiar las posiciones de las variables, utilice los botones de flecha arriba y abajo (Figura 18-39). Para realizar cambios en los factores intra-sujetos, puede volver a abrir el cuadro de diálogo *MLG Medidas repetidas: Definir factores* sin cerrar el cuadro de diálogo principal. Si lo desea, puede especificar covariables y factores inter-sujetos. En el fichero EMPLEADOS usaremos como factor intra-sujeto *minoría*, como variables dependientes *salario* y *tiememp*, como factores inter-sujetos *catlab* y *sexo* y como covariables *salini* y *exprev*. Los botones de la Figura 18-39 funcionan como en el MLG univariante y multivariante. Al pulsar *Aceptar* en la Figura 18-39 se obtiene la salida del procedimiento (Figuras 10-40 a 10-44).

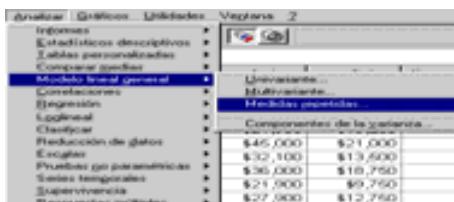


Figura 18-37

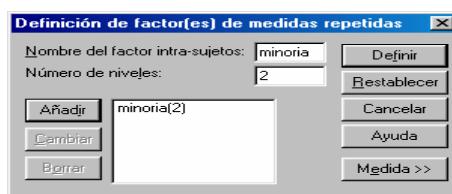


Figura 18-38

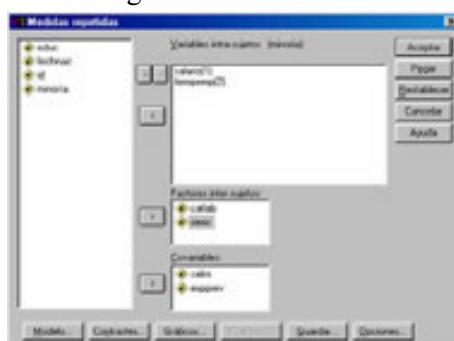


Figura 18-39

GLM salario tiememp BY catlab sexo WITH salini exprev /WEFACTOR=minoría 2 /EMMEANS=MEANS SSTYPE(3) /LINES PROFILE=catlab*sexo catlab*minoría sexo*minoría /CRITERIA = ALPHA(.05)/MSDESIGN = minoría /DESIGN = salini exprev catlab sexo catlab*sexo .																			
Factores intra-sujetos																			
Medida: MEASURE_1																			
<table border="1"> <thead> <tr> <th>MINORIA</th> <th>Variable dependiente</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>SALARIO</td> </tr> <tr> <td>2</td> <td>TIEMEMP</td> </tr> </tbody> </table>		MINORIA	Variable dependiente	1	SALARIO	2	TIEMEMP												
MINORIA	Variable dependiente																		
1	SALARIO																		
2	TIEMEMP																		
Factores inter-sujetos																			
<table border="1"> <thead> <tr> <th>Categoría</th> <th>Etiqueta del valor</th> <th>N</th> </tr> </thead> <tbody> <tr> <td>laboral</td> <td>Administrativo</td> <td>363</td> </tr> <tr> <td></td> <td>Seguridad</td> <td>27</td> </tr> <tr> <td></td> <td>Directivo</td> <td>84</td> </tr> <tr> <td>Sexo</td> <td>Hombre</td> <td>258</td> </tr> <tr> <td></td> <td>Mujer</td> <td>216</td> </tr> </tbody> </table>		Categoría	Etiqueta del valor	N	laboral	Administrativo	363		Seguridad	27		Directivo	84	Sexo	Hombre	258		Mujer	216
Categoría	Etiqueta del valor	N																	
laboral	Administrativo	363																	
	Seguridad	27																	
	Directivo	84																	
Sexo	Hombre	258																	
	Mujer	216																	

Figura 18-40

Contrastes multivariados ^b						
Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación
MINORIA	Traza de Pillai	,207	122,023 ^a	1,000	467,000	,000
	Lambda de Wilks	,793	122,023 ^a	1,000	467,000	,000
	Traza de Hotelling	,261	122,023 ^a	1,000	467,000	,000
	Raíz mayor de Roy	,261	122,023 ^a	1,000	467,000	,000
MINORIA * SALINI	Traza de Pillai	,429	350,564 ^a	1,000	467,000	,000
	Lambda de Wilks	,571	350,564 ^a	1,000	467,000	,000
	Traza de Hotelling	,751	350,564 ^a	1,000	467,000	,000
	Raíz mayor de Roy	,751	350,564 ^a	1,000	467,000	,000
MINORIA * EXPPREV	Traza de Pillai	,094	48,378 ^a	1,000	467,000	,000
	Lambda de Wilks	,906	48,378 ^a	1,000	467,000	,000
	Traza de Hotelling	,104	48,378 ^a	1,000	467,000	,000
	Raíz mayor de Roy	,104	48,378 ^a	1,000	467,000	,000
MINORIA * CATALAB	Traza de Pillai	,125	33,378 ^a	2,000	467,000	,000
	Lambda de Wilks	,875	33,378 ^a	2,000	467,000	,000
	Traza de Hotelling	,143	33,378 ^a	2,000	467,000	,000
	Raíz mayor de Roy	,143	33,378 ^a	2,000	467,000	,000
MINORIA * SEXO	Traza de Pillai	,012	5,550 ^a	1,000	467,000	,019
	Lambda de Wilks	,988	5,550 ^a	1,000	467,000	,019
	Traza de Hotelling	,012	5,550 ^a	1,000	467,000	,019
	Raíz mayor de Roy	,012	5,550 ^a	1,000	467,000	,019
MINORIA * CATALAB *	Traza de Pillai	,000	,172 ^a	1,000	467,000	,678
	Lambda de Wilks	,000	,172 ^a	1,000	467,000	,678
	Traza de Hotelling	,000	,172 ^a	1,000	467,000	,678
	Raíz mayor de Roy	,000	,172 ^a	1,000	467,000	,678

Figura 18-41

Pruebas de esfericidad de Mauchly ^b						
Medida: MEASURE_1	Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Epsilon ^a	
					Greenhouse o Geisser	Huynh Feldt
MINORIA		1,000	,000	0	,000	,000

Contraata la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

a. Puedo usarse para corregir los grados de libertad en las pruebas de significación promediadas. Las pruebas corregidas se muestran en la tabla Pruebas de los efectos intra-sujetos.

b. Diseño: Intercept+ SALINI+EXPPREV+CATLAB+SEXO+CATLAB * SEXO
Diseño intra sujetos: MINORIA

Figura 18-42

Pruebas de efectos intra-sujetos.						
Medida: MEASURE_1	Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
MINORIA	Esfericidad asumida	3031373150	1	3,03E+09	122,023	,000
	Greenhouse-Geisser	3031373150	1,000	3,03E+09	122,023	,000
	Huynh-Feldt	3031373150	1,000	3,03E+09	122,023	,000
	Límite-inferior	3031373150	1,000	3,03E+09	122,023	,000
MINORIA * SALINI	Esfericidad asumida	8708947694	1	8,71E+09	350,564	,000
	Greenhouse-Geisser	8708947694	1,000	8,71E+09	350,564	,000
	Huynh-Feldt	8708947694	1,000	8,71E+09	350,564	,000
	Límite-inferior	8708947694	1,000	8,71E+09	350,564	,000
MINORIA * EXPPREV	Esfericidad asumida	1201832425	1	1,20E+09	48,378	,000
	Greenhouse-Geisser	1201832425	1,000	1,20E+09	48,378	,000
	Huynh-Feldt	1201832425	1,000	1,20E+09	48,378	,000
	Límite-inferior	1201832425	1,000	1,20E+09	48,378	,000
MINORIA * CATALAB	Esfericidad asumida	1658404120	2	8,29E+08	33,378	,000
	Greenhouse-Geisser	1658404120	2,000	8,29E+08	33,378	,000
	Huynh-Feldt	1658404120	2,000	8,29E+08	33,378	,000
	Límite-inferior	1658404120	2,000	8,29E+08	33,378	,000
MINORIA * SEXO	Esfericidad asumida	137874775	1	1,38E+08	5,550	,019
	Greenhouse-Geisser	137874775	1,000	1,38E+08	5,550	,019
	Huynh-Feldt	137874775	1,000	1,38E+08	5,550	,019
	Límite-inferior	137874775	1,000	1,38E+08	5,550	,019
MINORIA * CATALAB *	Esfericidad asumida	4279272,462	1	4279272,5	,172	,678
	Greenhouse-Geisser	4279272,462	1,000	4279272,5	,172	,678
	Huynh-Feldt	4279272,462	1,000	4279272,5	,172	,678

Figura 18-43

Pruebas de contrastes intra-sujetos						
Medida: MEASURE_1						
Fuente	MINORIA	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
MINORIA	Lineal	3031373150	1	3,03E+09	122,023	,000
MINORIA * SALINI	Lineal	8708947694	1	8,71E+09	350,564	,000
MINORIA * EXPPREV	Lineal	1201832425	1	1,20E+09	48,378	,000
MINORIA * CATLAB	Lineal	1658404120	2	8,29E+08	33,378	,000
MINORIA * SEXO	Lineal	137874775	1	1,38E+08	5,550	,019
MINORIA * CATLAB *	Lineal	4279272,462	1	4279272,5	,172	,678
Error(MINORIA)	Lineal	1,160E+10	467	24842883		

Pruebas de los efectos inter-sujetos						
Medida: MEASURE_1						
Variable transformada: Promedio						
Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Intercept		3092885526	1	3,09E+09	124,342	,000
SALINI		8704442127	1	8,70E+09	349,941	,000
EXPPREV		1201757404	1	1,20E+09	48,314	,000
CATLAB		1658259923	2	8,29E+08	33,333	,000
SEXO		138647858	1	1,39E+08	5,574	,019
CATLAB * SEXO		4368262,445	1	4368262,4	,176	,675
Error		1,162E+10	467	24873993		

Figura 18-44

PROCEDIMIENTO COMPONENTES DE LA VARIANZA

El procedimiento Componentes de la varianza, para modelos de efectos mixtos, estima la contribución de cada efecto aleatorio a la varianza de la variable dependiente. Este procedimiento resulta de particular interés para el análisis de modelos mixtos, como los diseños *split-plot*, los diseños de medidas repetidas univariados y los diseños de bloques aleatorios. Al calcular las componentes de la varianza, se puede determinar dónde centrar la atención para reducir la varianza. Se dispone de cuatro métodos diferentes para estimar las componentes de la varianza: estimador mínimo no cuadrático insesgado (EMNCI, MINQUE), análisis de varianza (ANOVA), máxima verosimilitud (MV, ML) y máxima verosimilitud restringida (MVR, RML). Se dispone de diversas especificaciones para los diferentes métodos. Los resultados por defecto para todos los métodos incluyen las estimaciones de componentes de la varianza. Si se usa el método MV o el método MVR, se mostrará también una tabla con la matriz de covarianza asintótica. Otros resultados disponibles incluyen una tabla de ANOVA y las medias cuadráticas esperadas para el método ANOVA, y la historia de iteraciones para los métodos MV y MVR. El procedimiento Componentes de la varianza es totalmente compatible con el procedimiento MLG Factorial general. La opción Ponderación MCP permite especificar una variable usada para aplicar a las observaciones diferentes ponderaciones para un análisis ponderado; por ejemplo, para compensar la distinta precisión de las medidas.

Para obtener un análisis de componentes de la varianza, elija en los menús *Analizar* → *Modelo lineal general* → *Componentes de la varianza* (Figura 18-45), seleccione una variable dependiente y seleccione variables para Factor(es) fijo(s), Factor(es) aleatorio(s) y Covariable(s), en función de los datos (Figura 18-46). Para especificar una variable de ponderación, utilice *Ponderación MCP*. En el fichero EMPLEADOS usaremos como variable dependiente *salario*, como factores fijos *catlab* y *sexo*, como factor aleatorio minoría y como covariables *tiempemp* y *exprev*. Los botones *Modelo* (Figura 18-47) y *Método* (Figura 18-48) permiten elegir modelo y método de estimación. La salida se presenta en las Figuras 10-49 y 10-50.

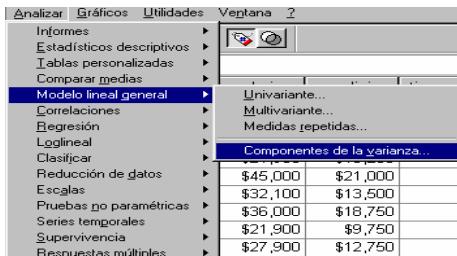


Figura 18-45



Figura 18-46

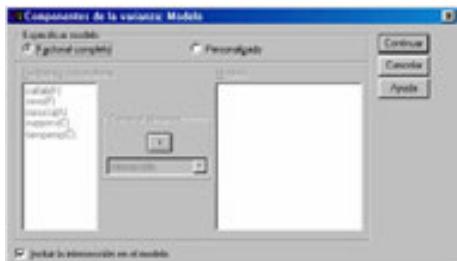


Figura 18-47

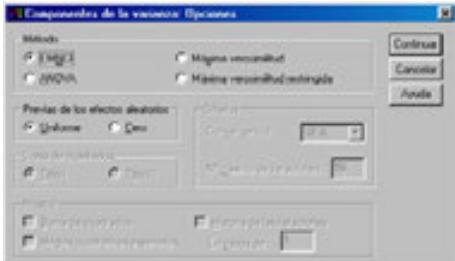


Figura 18-48

VARCOMP		
salario BY minoría catlab sexo WITH exprev tiempemp		
/RANDOM = minoría		
/METHOD = MINQUE (1)		
/DESIGN		
/INTERCEPT = INCLUDE .		
Estimación de componentes de la varianza		
Información sobre los niveles de los factores		
Clasificación	Etiqueta de Valor	
Clasificación	Sí	370
Étnica	2	104
Categoría laboral	Administrativo	363
	Seguridad	27
	Directivo	84
Sexo	Hombre	258
	Mujer	216

Figura 18-49

Estimaciones de la varianza	
Componente	Estimación
Var(MINORÍA)	-6969747,4 ^a
Var(MINORÍA * CATLAB)	22979979
Var(MINORÍA * SEXO)	-8519782,0 ^a
Var(MINORÍA * CATLAB * SEXO)	10530670
Var(Error)	83939164

Variable dependiente: SALARIO
Método: Estimación mínima no cuadrática insesgada
(Ponderación = 1 para Efectos aleatorios y Residual)

a. Con los métodos ANOVA y MINQUE pueden producirse estimaciones negativas de la componente de la varianza. Algunas razones para ello son: (a) el modelo especificado no es el correcto, o (b) el valor real de la varianza es igual a cero.

Figura 18-50

Ejercicio 18-1. Supongamos que disponemos de un conjunto de pinos clasificados por alturas (en metros) y por especies, según los datos siguientes:

Pinea	8,52	Pinaster	8,52	Pinea	8,13
Pinaster	6,45	Pinea	6,43	Halapensis	7,17
Silvestris	7,41	Pinea	6,21	Pinaster	8,40
Pinea	7,15	Halapensis	7,07	Silvestris	8,87
Pinaster	8,73	Pinaster	8,83	Pinea	6,12
Laricio	7,55	Pinaster	8,53	Pinaster	8,91
Halapensis	6,54	Laricio	7,84	Silvestris	8,81
Laricio	7,74	Silvestris	8,59	Laricio	7,40
Silvestris	8,65	Laricio	7,41	Pinaster	8,19
Silvestris	8,81	Pinaster	8,94	Pinaster	8,56

- a) Ajustar los datos a un modelo del análisis de la varianza para contrastar si pueden considerarse iguales todas las especies de pinos en cuanto a altura.
- b) Agrupar los datos en grupos homogéneos de especies en cuanto a altura realizando comparaciones de especies dos a dos.
- c) Hallar intervalos de confianza para las diferencias de alturas medias entre los diferentes pares de especies.

Comenzamos introduciendo los datos en dos variables (de nombres ESPECIE y ALTURA) en el editor de SPSS. Para la variable ESPECIE realizamos la codificación 1 = Pinea, 2 = Pinaster, 3 = Silvestris, 4 = Laricio y 5 = Halapensis. Para ajustar los datos a un modelo del análisis de la varianza consideraremos como variable respuesta la altura y como único factor la variable cualitativa especie cuyos niveles son 1, 2, 3, 4 y 5. Como tenemos un **modelo con un solo factor fijo**, utilizaremos el procedimiento ANOVA de un factor de la opción Comparar medias del menú Analizar, cuya pantalla de entrada se rellena como se indica en la Figura 18-51. Como el contraste a realizar es de igualdad de todas las medias se rellena la pantalla del botón *Contrastes* tal y como se indica en la Figura 18-52. La pantalla relativa al botón Post hoc se rellena como se indica en la Figura 18-53 con la finalidad de contrastar dos a dos las distintas especies en cuanto a altura media para formar grupos homogéneos de alturas, hallando a la vez intervalos de confianza para las diferencias de alturas medias (por 4 métodos). La pantalla del botón Opciones se rellena como se indica en la Figura 18-54 para contrastar la igualdad de varianzas entre las especies. La salida se muestra en las Figuras 10-55 a 10-60.

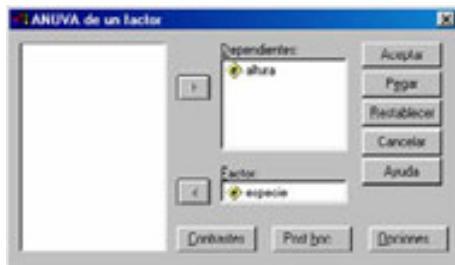


Figura 18-51

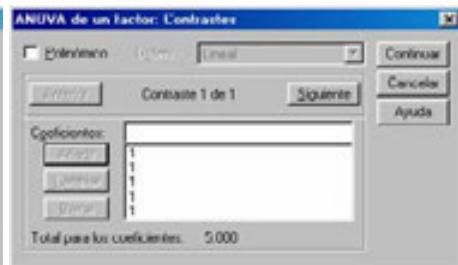


Figura 18-52

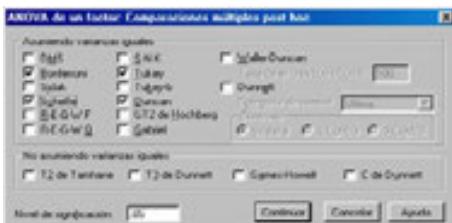


Figura 18-53

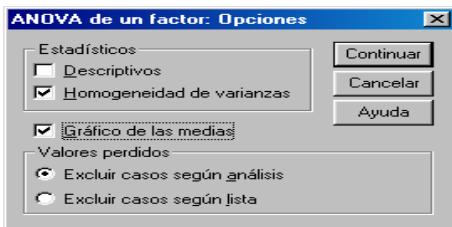


Figura 18-54

ONEWAY					
altura BY especie					
/CONTRASTES 1 1 1 1					
/STATISTICS HOMOGENEITY					
/PLOT MEANS					
/MISSING ANALYSIS					
/POSTHOC = TUKEY DUNCAN SCHEFFE BONFERRONI ALPHA(.05) .					
ANOVA de un factor					
Prueba de homogeneidad de varianzas					
ALTURA					
Estadístico de Levene	gl1	gl2	Sig.		
1,975	4	25	,129		
ANOVA					
ALTURA					
	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	12,116	4	3,029	6,332	,001
Intra-grupos	11,958	25	,478		
Total	24,074	29			

Figura 18-55

		ALTURA		
		Subconjunto para alfa = ,05		
ESPECIE		N	1	2
HSD de Tukey ^{a,b}	5	3	6,9267	
	1	6	,7,0933	
	4	5	,7,5880	7,5880
	2	10		8,4060
	3	6		8,5233
	Sig.		,549	,222
Duncan ^{a,b}	5	3	6,9267	
	1	6	,7,0933	
	4	5	,7,5880	7,5880
	2	10		8,4060
	3	6		8,5233
	Sig.		,158	,069
Scheffé ^{a,b}	5	3	6,9267	
	1	6	,7,0933	7,0933
	4	5	,7,5880	,7,5880
	2	10		8,4060
	3	6		8,5233
	Sig.		,672	,084

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 5,172.

b. Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

Pruebas para los contrastes

	Contraste	Valor del contraste	Error típico	t	gl	Sig. (bilateral)
ALTURA	Asumiendo igualdad de varianzas	1	38,5373 ^a	,6800	56,673	,25
	No asumiendo	1	38,5373 ^a	,5712	67,470	13,677

a. La suma de los coeficientes del contraste no es cero.

Figura 18-57

Comparaciones múltiples

		Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
(I) ESPECIE	(J) ESPECIE				Límite inferior	Límite superior
HSD de Tukey	1	-1,3127*	,3572	,009	-2,3616	,2637
	2	-1,4300*	,3993	,011	-2,6027	,2573
	3	-1,4947	,4188	,762	-1,7246	,7353
	4		,4890	,997	-1,2696	,1,6030
	5	,1667				
	2	1,3127*	,3572	,009	,2637	,2,3616
	3	-,1173	,3572	,997	-1,1663	,9316
	4	,8180	,3788	,228	-,2945	,1,9305
	5	1,4793*	,4553	,025	,1422	,2,8164
	3	1,4300*	,3993	,011	,2573	,2,6027
	2	,1173	,3572	,997	-,9316	,1,1663
	4	,9353	,4188	,200	-,2946	,2,1653
	5	1,5967*	,4890	,024	,1604	,3,0330
	4	,4947	,4188	,762	-,7353	,1,7246
	2	-,8180	,3788	,228	-1,9305	,2,945
	3	-,9353	,4188	,200	-2,1653	,2,946
	5	,6613	,5051	,688	-,8221	,2,1447
	5	-,1667	,4890	,997	-1,6030	,1,2696
	2	-,4793*	,4553	,025	-2,8164	,1,422
	3	-,15967*	,4890	,024	-3,0330	,1,604
	4	-,6613	,5051	,688	-2,1447	,8221

Figura 18-58

Scheffé	1	2	-1,3127*	,3572	,024	-2,4991	-,1263
	3		-1,4300*	,3993	,030	-2,7564	-,1036
	4		-,4947	,4188	,842	-1,8859	,8965
	5		,1667	,4890	,998	-1,4579	1,7912
	2	1	1,3127*	,3572	,024	,1263	2,4991
	3		-,1173	,3572	,998	-1,3037	1,0691
	4		,8180	,3788	,350	-4,404	2,0764
	5		1,4793	,4553	,058	-3,3049E-02	2,9917
	3	1	1,4300*	,3993	,030	,1036	2,7564
	2		,1173	,3572	,998	-1,0691	1,3037
	4		,9353	,4188	,317	-4,559	2,3265
	5		1,5967	,4890	,056	-2,7893E-02	3,2212
	4	1	,4947	,4188	,842	-,8965	1,8859
	2		-,8180	,3788	,350	-2,0764	,4404
	3		,9353	,4188	,317	-2,3265	,4559
	5		,6613	,5051	,787	-1,0165	2,3392
	5	1	-,1667	,4890	,998	-1,7912	1,4579
	2		-1,4793	,4553	,058	-2,9917	3,305E-02
	3		-1,5967	,4890	,056	-3,2212	2,789E-02
	4		-,6613	,5051	,787	-2,3392	1,0165

Figura 18-59

Bonferroni	1	2	-1,3127*	,3572	,011	-2,4120	-,2133
	3		-1,4300*	,3993	,014	-2,6591	-,2009
	4		-,4947	,4188	,1000	-1,7838	,7945
	5		,1667	,4890	,1000	-1,3387	1,6721
	2	1	1,3127*	,3572	,011	,2133	2,4120
	3		-,1173	,3572	,1000	-1,2167	,9820
	4		,8180	,3788	,406	-,3481	1,9841
	5		1,4793*	,4553	,033	7,789E-02	2,8808
	3	1	1,4300*	,3993	,014	,2009	2,6591
	2		,1173	,3572	,1000	-,9820	1,2167
	4		,9353	,4188	,347	-,3538	2,2245
	5		1,5967*	,4890	,032	9,128E-02	3,1021
	4	1	,4947	,4188	,1000	-,7945	1,7838
	2		-,8180	,3788	,406	-1,9841	,3481
	3		,9353	,4188	,347	-2,2245	,3538
	5		,6613	,5051	,1000	-,8934	2,2161
	5	1	-,1667	,4890	,1000	-1,6721	1,3387
	2		-1,4793*	,4553	,033	-2,8808	-7,79E-02
	3		-1,5967*	,4890	,032	-3,1021	-9,13E-02
	4		-,6613	,5051	,1000	-2,2161	,8934

Figura 18-60

La prueba de homogeneidad de varianzas de la Figura 18-55 presenta un p-valor mayor que 0,05, con lo que se acepta la igualdad de varianzas para los distintos niveles (especies). En cuanto a grupos homogéneos (Figura 18-56), según la prueba HSD de Tukey forman un grupo homogéneo en cuanto a altura media las especies 5, 1 y 4, o sea, las especies Halapensis, Pinea y Laricio. También forman otro grupo homogéneo las especies 2, 3 y 4, o sea, las especies Pinaster, Silvestris y Laricio. Según la prueba de Dunkan el primer grupo homogéneo coincide con el anterior, el segundo grupo homogéneo lo forman las especies 2 y 4, y un tercer grupo lo forman las especies 2 y 3. Según la prueba de Scheffé el primer grupo homogéneo sigue siendo el mismo, el segundo lo forman las especies 1, 2 y 4 y el tercero lo forman las especies 2, 3 y 4. Por otra parte, como el p-valor del test de F de Fisher Snedokor de igualdad de todas las medias de los niveles es menor que 0,05 (Figura 18-57), existen diferencias significativas entre las alturas medias de los pinos de las diferentes especies al 95% de confianza.

En las Figuras 10-58 a 10-60 y su columna *Diferencia de medias* se observan, señalados con un asterisco, los pares de clases de pinos que presentan diferencias significativas en cuanto a altura. La columna *Intervalos de confianza al 95%* define los intervalos de confianza para las diferencias de medias. Se observa que los intervalos de confianza que no contienen el cero, son los que aparecen marcados en la columna *Diferencia de medias* con un asterisco que indica diferencias significativas entre las medias de dichos niveles, es decir, entre las alturas medias de las especies.

Ejercicio 18-2. *Se sospecha que hay variabilidad para la preparación del examen de selectividad, entre los diferentes centros de bachillerato de una región. Con el fin de estudiarla, se eligieron 5 centros al azar de entre todos los centros de la región, cuyo número es muy grande. De cada centro seleccionado se eligieron 8 alumnos al azar, con la condición de que hubieran cursado las mismas asignaturas, y se anotaron las calificaciones que obtuvieron en el examen de selectividad. Los resultados fueron:*

	1 →	5,5	5,2	5,9	7,1	6,2	5,9	5,3	6,2
	2 →	6,1	7,2	5,5	6,7	7,6	5,9	8,1	8,3
<i>Centros</i>	3 →	4,9	5,5	6,1	6,1	6,2	6,4	6,9	4,5
	4 →	3,2	3,3	5,5	5,7	6,0	6,1	4,7	5,1
	5 →	6,7	5,8	5,4	5,5	4,9	6,2	6,1	7,0

Estudiar si existe variabilidad entre centros y estimar esta variabilidad. ¿Qué centros son los mejores en la preparación de la selectividad?

Como los centros han sido elegidos al azar dentro de una población total de centros que se supone muy grande, interpretaremos que los datos se adaptan a un **diseño de un factor de efectos aleatorios**, es decir, estamos ante un **modelo de componentes de la varianza unifactorial**. Comenzamos introduciendo los datos en dos variables (de nombres CENTRO y CALIFI) en el editor de SPSS.

Para ajustar los datos a un modelo del análisis de la varianza consideramos como variable respuesta las calificaciones y como único factor la variable cualitativa centro cuyos niveles son 1, 2, 3, 4 y 5. Como tenemos un modelo con un solo factor aleatorio, utilizaremos el procedimiento *Modelo Lineal General Univariante*, cuya pantalla de entrada se rellena como se indica en la Figura 18-61. Las pantallas de los botones *Opciones* y *Contrastes* se llenan como se indica en las Figuras 10-62 y 10-63. Parte de la salida se presenta en las Figuras 10-64 (p valor mayor que 0,05 que lleva a aceptar la igualdad de varianzas de los grupos), 10-65 (p-valor del test de F menor que 0,05 con lo que existen diferencias significativas entre las calificaciones de los diferentes centros al 95% de confianza y hay que admitir que existe una variabilidad entre centros) y 10-66 (los únicos centros con diferencias significativas son el 2 y el 4).

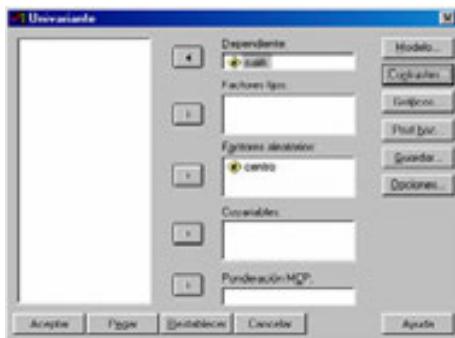


Figura 18-61

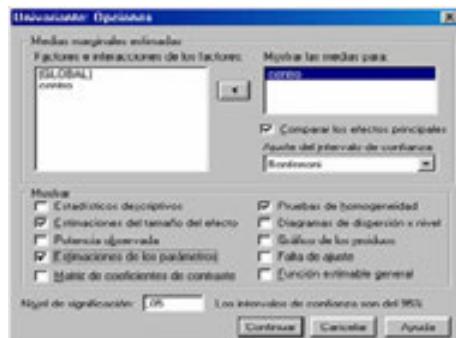


Figura 18-62



Figura 18-63

Contraste de Levene sobre la igualdad de las varianzas error ^a						
Variable dependiente: CALIFI						
	F	gl1	gl2	Significación		
1,628		4	35		,189	

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+CENTRO

Figura 18-64

Resultados de la prueba

Variable dependiente: CALIFI

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Significación	Eta cuadrado
Contraste	15,685	4	3,921	5,031	,003	,365
Error	27,279	35	,779			

Figura 18-65

Comparaciones por pares						
Variable dependiente: CALIFI						
(I) CENTRO	(J) CENTRO	Diferencia entre medias (I-J)	Error tip.	Significación ^a	Intervalo de confianza al 95 % para diferencia ^a	
					Límite inferior	Límite superior
1	2	-1,013	,441	,279	-2,335	,310
	3	8,750E-02	,441	1,000	-1,235	1,410
	4	,963	,441	,360	-,360	2,285
	5	-3,750E-02	,441	1,000	-1,360	1,285
	2					
2	1	1,013	,441	,279	-,310	2,335
	3	1,100	,441	,176	-,223	2,423
	4	1,975*	,441	,001	,652	3,298
	5	,875	,441	,553	-,448	2,198
	3	-,125	,441	1,000	-1,448	1,198
3	1	-8,750E-02	,441	1,000	-1,410	1,235
	2	-,1000	,441	,176	-2,423	,223
	4	,875	,441	,553	-,448	2,198
	5	-,125	,441	,298	-2,323	,323
	4					
4	1	-,963	,441	,360	-,285	,360
	2	-1,975*	,441	,001	-3,298	-,652
	3	-,875	,441	,553	-2,198	,448
	5	-,1000	,441	,298	-2,323	,323
	5					
Basadas en las medias marginales estimadas.						
*. La diferencia de las medias es significativa al nivel ,05.						

Figura 18-66

Para estimar la variabilidad entre centros utilizamos el procedimiento *Componentes de la varianza* de la opción *Modelo Lineal General* del menú *Analizar*, que nos ofrece las estimaciones de las componentes de la varianza. Rellenamos su pantalla de entrada como se indica en la Figura 18-67. La pantalla del botón *Opciones* se rellena como se indica en la Figura 18-68. La Figura 18-69 presenta la tabla ANOVA y la Figura 18-70 presenta las estimaciones de la variabilidad. Se observa que la estimación para la varianza del error es 0,779, y la estimación para la varianza del factor es 0,393. Luego la variabilidad entre centros se estima en 0,393, o en su raíz cuadrada en caso de medir la variabilidad a través de la desviación típica.



Figura 18-67

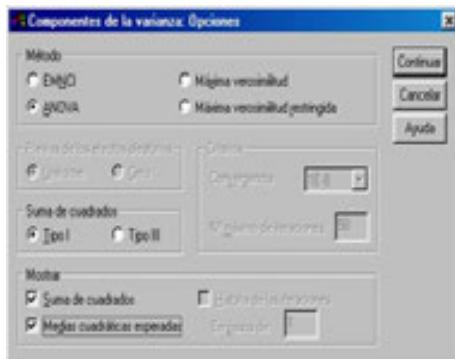


Figura 18-68

ANOVA			
Fuente	Suma de cuadrados tipo I	gl	Media cuadrática
Modelo corregido	15,685	4	3,921
Intersección	1398,306	1	1398,306
CENTRO	15,685	4	3,921
Error	27,279	35	.779
Total	1441,270	40	
Total corregido	42,964	39	

Variáble dependiente: CALIFI

Figura 18-69

Estimaciones de la varianza	
Componente	Estimación
Var(CENTRO)	,393
Var/Error)	.779

Variáble dependiente: CALIFI
Método: ANOVA (Tipo I Suma de cuadrados)

Figura 18-70

Ejercicio 18-3. Ajustar un modelo de análisis de la covarianza que hace depender la variable consumo (consumo de gasolina de los coches) de los factores categóricos año (año de fabricación) y origen (origen geográfico) y de los covariantes o factores cuantitativos cv (potencia en caballos) y peso (peso).

Comenzamos rellenando la pantalla de entrada del procedimiento *Modelo Lineal General Univariante* como se indica en la Figura 18-71. Las pantallas de los botones *Modelo* y *Opciones* se llenan como se indica en las Figuras 10-72 y 10-73, ya que inicialmente consideramos la estimación del modelo completo. Al pulsar *Aceptar* se obtiene la prueba de efectos inter-sujetos de la Figura 18-74 que indica que son significativos al 95% todos los factores (p-valores menores que 0,05), pero no lo es la interacción del año y el origen.

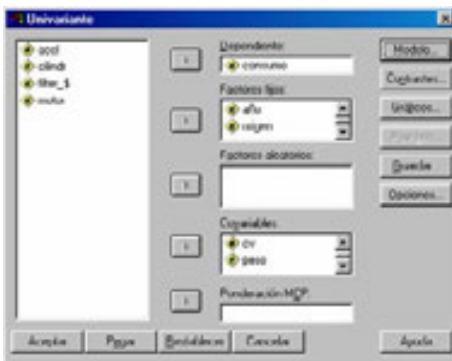


Figura 18-71



Figura 18-72

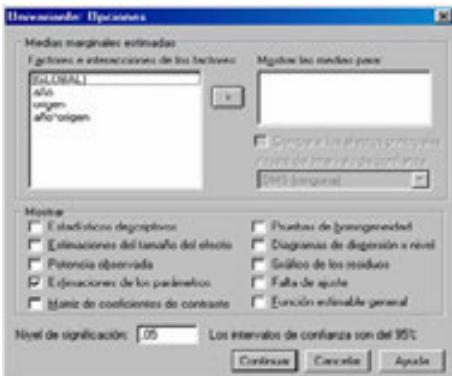


Figura 18-73

Pruebas de los efectos inter-sujetos					
	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	5339,009 ^a	40	133,475	81,710	,000
Intersección	22,026	1	22,026	13,484	,000
CV	29,449	1	29,449	18,028	,000
PESO	317,541	1	317,541	194,390	,000
ANO	306,929	12	25,577	15,656	,000
ORIGEN	18,720	2	9,360	5,730	,004
ANO * ORIGEN	40,550	24	1,690	1,034	,421
Error	571,733	350			
Total	55200,000	391			
Total corregida	5910,742	390			

^a R cuadrado = ,903 (R cuadrado corregida = ,892)

Figura 18-74

Para realizar la estimación definitiva del modelo, y dado que la interacción entre año y origen no es significativa, rellenaremos la pantalla del botón *Modelo* como se indica en la Figura 18-75. La pantalla del botón *Opciones* se rellena como se indica en la Figura 18-76. Al pulsar *Aceptar* se obtiene la estimación del modelo (Figura 18-77), la prueba de ajuste de la Figura 18-78 con p-valor mayor de 0,05 (ajuste correcto) y la prueba de igualdad de varianzas para los grupos de la Figura 18-79 con p-valor mayor que 0,05 (se acepta la igualdad de varianzas).

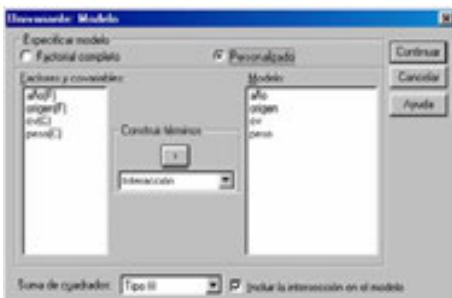


Figura 18-75

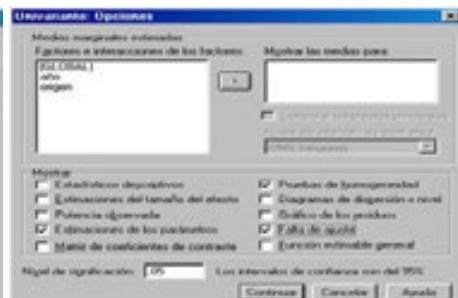


Figura 18-76

Estimaciones de los parámetros

Variable dependiente: Consumo (l/100Km)

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Intersección	-,735	,319	-2,304	,022	-1,363	-,108
[AÑO=70]	3,010	,371	8,114	,000	2,281	3,739
[AÑO=71]	2,591	,345	7,513	,000	1,913	3,270
[AÑO=72]	3,155	,348	9,054	,000	2,470	3,840
[AÑO=73]	3,566	,328	10,874	,000	2,922	4,211
[AÑO=74]	2,386	,348	6,846	,000	1,701	3,071
[AÑO=75]	2,317	,341	6,801	,000	1,847	2,987
[AÑO=76]	2,004	,329	6,089	,000	1,357	2,651
[AÑO=77]	1,374	,341	4,023	,000	,702	2,045
[AÑO=78]	1,432	,320	4,470	,000	,802	2,061
[AÑO=79]	,377	,339	1,114	,266	-,289	1,043
[AÑO=80]	-,142	,348	-,410	,682	-,826	,541
[AÑO=81]	,334	,339	,986	,325	-,332	1,001
[AÑO=82]	0 ^a	,	,	,	,	,
[ORIGEN=1]	,376	,207	1,813	,071	-3,171E-02	,783
[ORIGEN=2]	-,253	,220	-1,148	,252	-,685	,180
[ORIGEN=3]	0 ^a	,	,	,	,	,
CV	1,985E-02	,004	5,009	,000	1,206E-02	2,764E-02
PESO	8,010E-03	,001	14,624	,000	6,933E-03	9,087E-03

a. Al parámetro se le ha asignado el valor cero porque es redundante.

Figura 18-77

Pruebas de falta de ajuste					
Variable dependiente: Consumo (l/100Km)					
Fuente	Suma de cuadrados	gl	Media cuadrática	F	Significación
Falta de ajuste	611,282	372	1,643	3,286	,262
Error puro	1,000	2	,500		

Figura 18-78

Contraste de Levene sobre la igualdad de las varianzas error^a

Variable dependiente: Consumo (l/100Km)

F	gl1	gl2	Significación
1,364	38	352	,081

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+AÑO+ORIGEN+CV+PESO

Figura 18-79

Ejercicio 18-4. Se desea estudiar la eficacia de dos somníferos A_1 y A_2 frente a un placebo, teniendo en cuenta la posible influencia del sexo B_1 , B_2 y la forma de administración C_1 (intramuscular), C_2 (oral) y C_3 (intravenosa). Se tomaron 3 réplicas para cada una de las combinaciones de los niveles de los factores, obteniéndose los siguientes datos de eficacia en horas de sueño.

		A_1		A_2		A_3	
		B_1	B_2	B_1	B_2	B_1	B_2
C_1	7,3	7,6		8,5	8,3	6,7	6,1
	7,5	7,4		8,3	8,7	6,5	6,2
	7,1	7,2		8,4	7,9	6,3	6,9
C_2	7,1	6,8		7,5	7,6	6,7	6,4
	7,3	7,3		7,2	7,4	6,3	6,9
	6,9	7,2		7,2	7,2	6,2	6,8
C_3	8,1	8,3		8,9	9,0	6,8	6,0
	8,2	8,2		8,4	8,5	6,2	6,1
	8,0	8,1		8,1	8,0	6,2	6,2

a) Analizar la significación de los efectos principales y de las interacciones.

b) Representar gráficamente las interacciones.

Comenzamos introduciendo los datos en cuatro variables de nombres TRATAM (tipo de somníferos), SEXO, FORMA (forma de administración) y EFICAC (eficacia de los somníferos) en el editor de SPSS. La variable TRATAM se codifica mediante 1 = A₁, 2 = A₂ y 3 = A₃. La variable SEXO se codifica como 1 = B₁ y 2 = B₂. La variable FORMA se codifica como 1 = C₁, 2 = C₂ y 3 = C₃.

Tenemos como causas de la variabilidad de la variable respuesta eficacia (EFICAC) los tres factores tipo de somnífero (TRATAM), sexo (SEXO) y forma de administración (FORMA), con 3, 2 y 3 niveles cada uno, y con 3 réplicas por casilla. Estaremos entonces ante un **diseño de tres factores con interacción replicado**. Por lo tanto, rellenamos la pantalla de entrada del procedimiento *Modelo Lineal General Univariante* como se indica en la Figura 18-80. Las pantallas de los botones *Opciones*, *Gráficos* y *Contrastes* se llenan como se indica en las Figuras 10-81 a 10-83. Al pulsar *Aceptar* se obtiene el contraste de igualdad de varianzas de grupos (se acepta porque el p-valor es mayor que 0,5) y la prueba de efectos inter-sujetos de la Figura 18-84, que indica que son significativos al 95% los efectos principales tratamiento y forma así o como su interacción (p-valores menores que 0,05), pero no lo son ni el efecto principal sexo ni el resto de las interacciones. Los gráficos de interacción de las Figuras 10-85 a 10-87 confirman que sólo es significativa la interacción de tratamiento y forma (es el único gráfico cuyas líneas se cruzan).

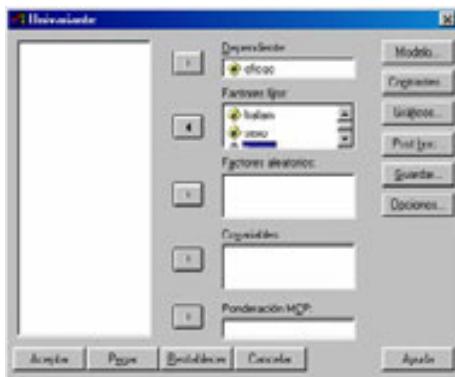


Figura 18-80

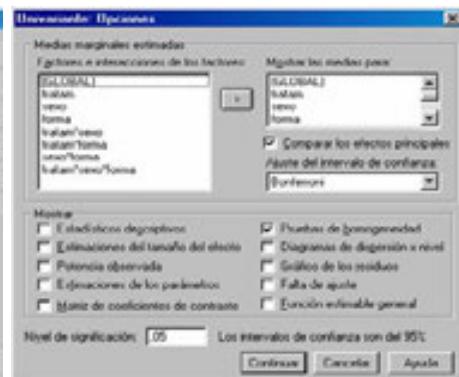


Figura 18-81



Figura 18-82



Figura 18-83

Contraste de Levene sobre la igualdad de las varianzas error ^a					
Variable dependiente: EFICAC	F	gl1	gl2	Significación	
	1,361	17	36		,213

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+TRATAM+SEXO+FORMA+TRATAM * SEXO+TRATAM * FORMA+SEXO * FORMA+TRATAM * SEXO * FORMA

Pruebas de los efectos inter-sujetos						
Variable dependiente: EFICAC	Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido		34,219 ^a	17	2,013	26,576	,000
Intercept		2906,934	1	2906,934	38380,059	,000
TRATAM		25,378	2	12,689	167,533	,000
SEXO		2,963E-03	1	2,963E-03	,039	,844
FORMA		3,605	2	1,802	23,797	,000
TRATAM * SEXO		2,259E-02	2	1,130E-02	,149	,867
TRATAM * FORMA		4,890	4	1,222	16,139	,000
SEXO * FORMA		9,593E-02	2	4,798E-02	,633	,537
TRATAM * SEXO * FORMA		,225	4	5,630E-02	,743	,569
Error		2,727	36			
Total		2943,880	54			
Total corregida		36,946	53			

a. R cuadrado = ,926 (R cuadrado corregida = ,891)

Figura 18-84

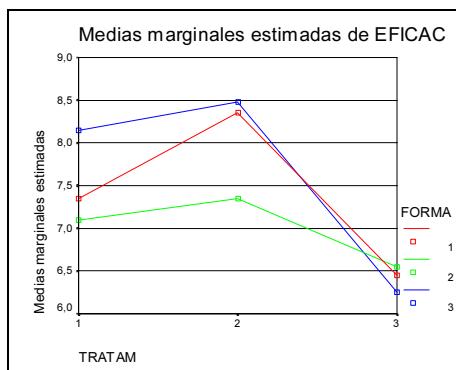


Figura 18-85

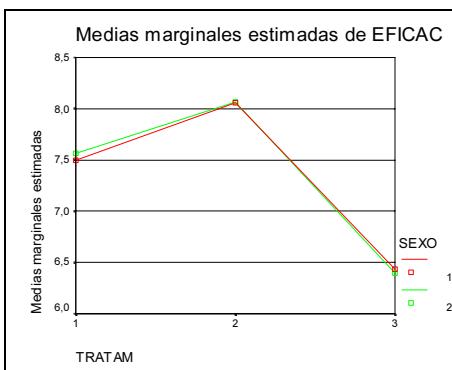


Figura 18-86

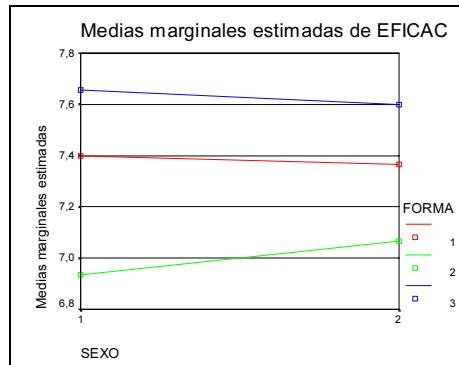


Figura 18-87

Ejercicio 18-5. Un fabricante de plásticos mide tres propiedades de la película de plástico: resistencia, brillo y opacidad. Se prueban dos tasas de extrusión y dos cantidades diferentes de aditivo y se miden las tres propiedades para cada combinación de tasa de extrusión y cantidad de aditivo obteniéndose los siguientes resultados:

estru	aditivo	resist	brillo	opacidad	estru	aditivo	resist	brillo	opacidad
1 1		6,5	9,5	4,4	2		6,7	9,1	2,8
1 1		6,2	9,9	6,4	2	1	6,6	9,3	4,1
1 1		5,8	9,6	3,0	2	1	7,2	8,3	3,8
1 1		6,5	9,6	4,1	2	1	7,1	8,4	1,6
1 1		6,5	9,2	0,8	2	1	6,8	8,5	3,4
1 2		6,9	9,1	5,7	2	2	7,1	9,2	8,4
1 2		7,2	10,0	2,0	2	2	7,0	8,8	5,2
1 2		6,9	9,9	3,9	2	2	7,2	9,7	6,9
1 2		6,1	9,5	1,9	2	2	7,5	10,1	2,7
1 2		6,3	9,4	5,7	2	2	7,6	9,2	1,9

El fabricante quiere saber si la tasa de extrusión y la cantidad de aditivo producen individualmente resultados significativos. También quiere saber si la interacción de los dos factores no es significativa.

Tenemos tres variables dependientes y un factor fijo, por lo que rellenaremos la pantalla de entrada del procedimiento *Modelo Lineal General Multivariante* como se indica en la Figura 18-88. Las pantallas de los botones *Opciones*, *Contrastes* y *Gráficos* se llenan como se indica en las Figuras 10-89 a 10-91. Al pulsar *Aceptar* se obtiene la tabla de contrastes multivariados de la Figura 18-92 que indica que la tasa de extrusión y la cantidad de aditivo tienen efectos significativos al 95% (p-valores menores que 0,05), mientras que el efecto de la interacción no es significativo (p-valor mayor que 0,05). Por otra parte, la prueba de efectos inter-sujetos de la Figura 18-93 indica que son significativos al 95% los efectos principales cantidad de aditivo y tasa de extrusión, así como su interacción (p-valores menores que 0,05) para los factores resistencia a la ruptura y brillo, pero no lo son para el factor opacidad (p-valor mayor que 0,05). Los gráficos de interacción de las Figuras 10-94 a 10-96 confirman que la tasa de extrusión y la cantidad de aditivo sólo interaccionan para el factor opacidad (es el único caso en que las líneas se cruzan).

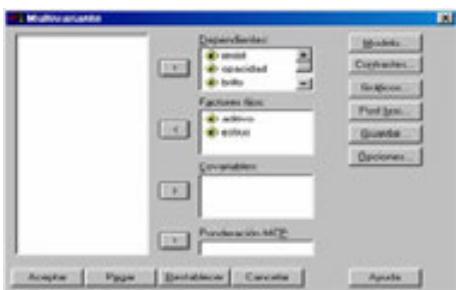


Figura 18-88



Figura 18-89



Figura 18-90



Figura 18-91

Contrastes multivariados ^b					
Efecto	Valor	F	Gl de la hipótesis	Gl del error	Significación
Intercept	Traza de Pillai	,999	5950,906 ^a	3,000	14,000 ,000
	Lambda de Wilks	,001	5950,906 ^a	3,000	14,000 ,000
	Traza de Hotelling	1275,194	5950,906 ^a	3,000	14,000 ,000
	Raíz mayor de Roy	1275,194	5950,906 ^a	3,000	14,000 ,000
ADITIVO	Traza de Pillai	,477	4,256 ^a	3,000	14,000 ,025
	Lambda de Wilks	,523	4,256 ^a	3,000	14,000 ,025
	Traza de Hotelling	,912	4,256 ^a	3,000	14,000 ,025
	Raíz mayor de Roy	,912	4,256 ^a	3,000	14,000 ,025
ESTRUS	Traza de Pillai	,618	7,554 ^a	3,000	14,000 ,003
	Lambda de Wilks	,382	7,554 ^a	3,000	14,000 ,003
	Traza de Hotelling	1,619	7,554 ^a	3,000	14,000 ,003
	Raíz mayor de Roy	1,619	7,554 ^a	3,000	14,000 ,003
ADITIVO * ESTRUS	Traza de Pillai	,223	1,339 ^a	3,000	14,000 ,302
	Lambda de Wilks	,777	1,339 ^a	3,000	14,000 ,302
	Traza de Hotelling	,287	1,339 ^a	3,000	14,000 ,302
	Raíz mayor de Roy	,287	1,339 ^a	3,000	14,000 ,302

Figura 18-92

Pruebas de los efectos inter-sujetos

Fuente	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	Resistencia a la ruptura	2,501 ^a	3	,834	7,563	,002
	Opacidad	9,282 ^b	3	3,094	,762	,531
	Brillo	2,457 ^c	3	,819	4,987	,012
Intercept	Resistencia a la ruptura	920,724	1	920,724	8351,243	,000
	Opacidad	309,684	1	309,684	76,319	,000
	Brillo	1735,384	1	1735,384	10565,507	,000
ADITIVO	Resistencia a la ruptura	,760	1	,760	6,898	,018
	Opacidad	4,900	1	4,900	1,208	,288
	Brillo	,612	1	,612	3,729	,071
ESTRUS	Resistencia a la ruptura	1,740	1	1,740	15,787	,001
	Opacidad	,421	1	,421	,104	,752
	Brillo	1,301	1	1,301	7,918	,012
ADITIVO * ESTRUS	Resistencia a la ruptura	5,000E-04	1	5,000E-04	,005	,947
	Opacidad	3,960	1	3,960	,976	,338
	Brillo	,544	1	,544	3,315	,087
Error	Resistencia a la ruptura	1,764	16	,110		
	Opacidad	64,924	16	4,058		
	Brillo	2,628	16	,164		
Total	Resistencia a la ruptura	924,990	20			
	Opacidad	383,890	20			
	Brillo	1740,470	20			
Total corregida	Resistencia a la ruptura	4,265	19			
	Opacidad	74,206	19			
	Brillo	5,085	19			

Figura 18-93

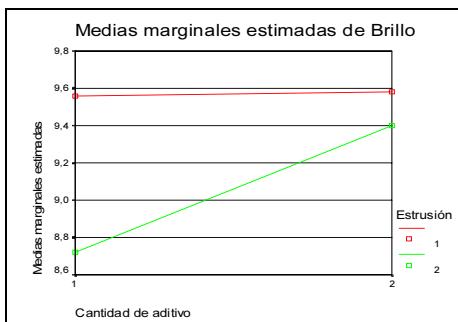


Figura 18-94

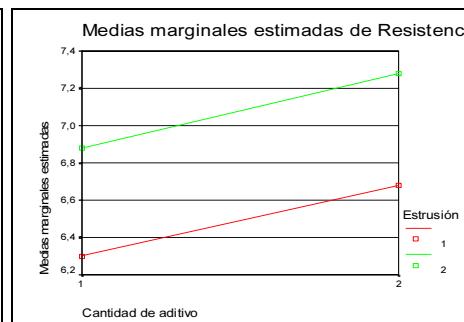


Figura 18-95

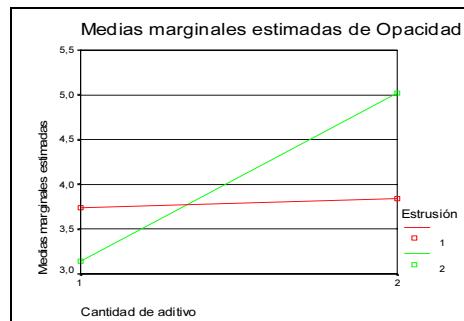


Figura 18-96

Ejercicio 18-6. Se asignan doce estudiantes a un grupo de alta o de baja ansiedad ante una tarea de aprendizaje basándose en las puntuaciones obtenidas en una prueba de nivel de ansiedad. El nivel de ansiedad es un factor inter-sujetos que divide a los sujetos en grupos. A cada estudiante se le dan cuatro ensayos para una tarea de aprendizaje y se registra el número de errores por ensayo. Los errores de cada ensayo se registran en variables distintas y se define un factor intra-sujetos (ensayo) con cuatro niveles para cada uno de los ensayos. Los datos son:

	Sujeto	Ansiedad	Tensión	Ensay1	Ensay2	Ensay3	Ensay4
1	1	1	1	18	14	12	6
2	1	1	1	19	12	8	4
3	1	1	1	14	10	6	2
4	1	2	2	16	12	10	4
5	1	2	2	12	8	6	2
6	1	2	2	18	10	5	1
7	2	1	1	16	10	8	4
8	2	1	1	18	8	4	1
9	2	1	1	16	12	6	2
10	2	2	2	19	16	10	8
11	2	2	2	16	14	10	9
12	2	2	2	16	12	8	8

Se trata de descubrir si el efecto de los ensayos es significativo y si la interacción ensayo-ansiedad es o no significativa.

Comenzamos rellenando la pantalla de entrada del procedimiento *Modelo Lineal General en Medidas Repetidas* como se indica en la Figura 18-97. Se pulsa *Añadir* y *Definir* y se rellena la pantalla *Medidas repetidas* como se indica en la Figura 18-98. Las pantallas de los botones *Contrastes*, *Gráficos* y *Opciones* se rellenan como se indica en las Figuras 10-99 a 10-101. Al pulsar *Aceptar* se obtiene la prueba de igualdad de covarianzas en los grupos (se acepta la igualdad porque el p-valor es mayor que 0,05) de la Figura 18-102, la tabla de contrastes multivariados de la Figura 18-103 que indica que la significatividad del efecto ensayo (p-valor menor que 0,05) y la no significatividad de la interacción ensayo-ansiedad (p-valor mayor que 0,05). Por otra parte, las pruebas de efectos inter-sujetos y de contrastes intra-sujetos de las Figuras 10-104 y 10-105 corroboran la significatividad del efecto ensayo y la no significatividad de la interacción ensayo-ansiedad. En el gráfico de perfil de la Figura 18-106 se observa que las ansiedades para los cuatro ensayos no interaccionan.

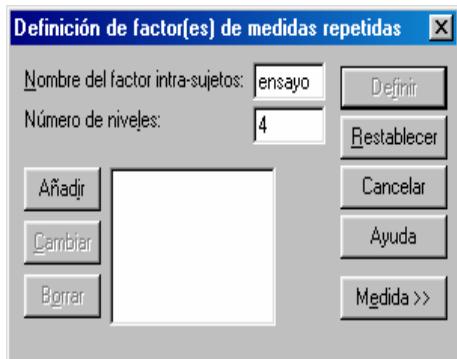


Figura 18-97



Figura 18-98



Figura 18-99

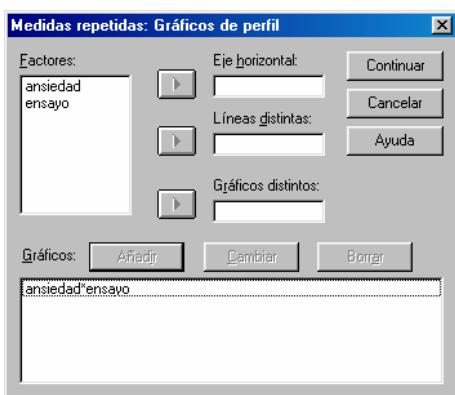


Figura 18-100

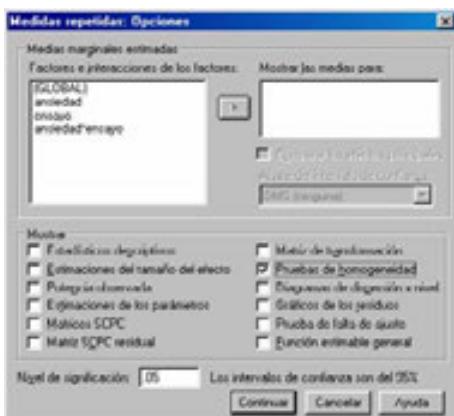


Figura 18-101

Factores inter-sujetos		N
Ansiedad	1	6
	2	6

Prueba de Box sobre la igualdad de las matrices de covarianza

M de Box	21,146
F	1,161
gl1	10
gl2	478,088
Significación	,315

Contrasta la hipótesis nula de que las matrices de covarianza observadas de las variables dependientes son iguales en todos los grupos.

a.

Diseño: Intercept+ANSIEDAD
Diseño intra sujetos: ENSAYO

Figura 18-102

Contrastes multivariados*						
Efecto	Valor	F	Ot de la hipótesis	Ot del error	Significación	
ENSAYO	Trazo de Pillai	,961	64,054*	3,000	0,000	,000
	Lambda de Wilks	,039	64,054*	3,000	0,000	,000
	Trazo de Hotelling	24,320	64,054*	3,000	0,000	,000
	Raíz mayor de Roy	24,320	64,054*	3,000	0,000	,000
ENSAYO * ANSIEDAD	Trazo de Pillai	,479	2,451*	3,000	0,000	,130
	Lambda de Wilks	,521	2,451*	3,000	0,000	,138
	Trazo de Hotelling	,919	2,451*	3,000	0,000	,138
	Raíz mayor de Roy	,919	2,451*	3,000	0,000	,138

* Estadístico exacto
b.

Diseño: Intercept+ANSIEDAD
Diseño intra sujetos: ENSAYO

Prueba de esfericidad de Mauchly^b

Efecto intra sujetos	W de Mauchly	Cfr-cuadrado aprox.	gl	Significación	Epsilon ^a		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
ENSAYO	,284	11,011	5	,002	,844	,701	,724

Contrasta la hipótesis nula de que la matriz de covariancia error de las variables dependientes transformadas es proporcional a una matriz identidad.

Figura 18-103

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
ENSAYO	Esfericidad asumida	991,500	3	330,500	128,627 ,000
	Greenhouse-Geisser	991,500	1,632	607,468	128,627 ,000
	Huynh-Feldt	991,500	2,102	471,773	128,627 ,000
	Límite-inferior	991,500	1,000	991,500	128,627 ,000
ENSAYO * ANSIEDAD	Esfericidad asumida	8,417	3	2,806	1,092 ,368
	Greenhouse-Geisser	8,417	1,632	5,157	1,092 ,346
	Huynh-Feldt	8,417	2,102	4,005	1,092 ,357
	Límite-inferior	8,417	1,000	8,417	1,092 ,321
Error(ENSAYO)	Esfericidad asumida	77,083	30	2,569	
	Greenhouse-Geisser	77,083	16,322	4,723	
	Huynh-Feldt	77,083	21,016	3,668	
	Límite-inferior	77,083	10,000	7,708	

Figura 18-104

Pruebas de contrastes intra-sujetos							
Medida: MEASURE_1	Fuente	ENSAYO	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
ENSAYO	Nivel 1 - Nivel 4	1800,750	1	1800,750	177,414	,000	
	Nivel 2 - Nivel 4	630,750	1	630,750	223,935	,000	
	Nivel 3 - Nivel 4	147,000	1	147,000	78,750	,000	
ENSAYO * ANSIEDAD	Nivel 1 - Nivel 4	6,750	1	6,750	,665	,434	
	Nivel 2 - Nivel 4	4,083	1	4,083	1,450	,256	
	Nivel 3 - Nivel 4	16,333	1	16,333	8,750	,014	
Error(ENSAYO)	Nivel 1 - Nivel 4	101,500	10	10,150			
	Nivel 2 - Nivel 4	28,167	10	2,817			
	Nivel 3 - Nivel 4	18,667	10	1,867			

Contraste de Levene sobre la igualdad de las varianzas error ^a				
	F	gl1	gl2	Significación
Ensayo 1	3,312	1	10	,099
Ensayo 2	,156	1	10	,701
Ensayo 3	,266	1	10	,617
Ensayo 4	7,788	1	10	,019

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

Figura 18-105

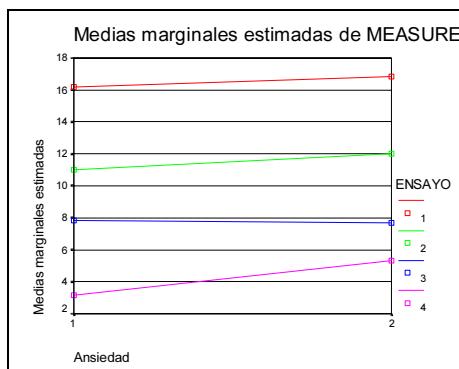


Figura 18-106

MODELOS DE ELECCIÓN DISCRETA LOGIT Y PROBIT. REGRESIÓN DE COX

MODELOS CON VARIABLES CUALITATIVAS: MODELOS DE ELECCIÓN DISCRETA

Aunque el rango de valores posibles de una variable estadística es usualmente un intervalo de la recta real, sin embargo la teoría econométrica considera también modelos de regresión lineal en los que alguna de las variables explicativas toma valores en un conjunto discreto y finito. Un ejemplo de este tipo de variables lo constituye las llamadas *variables ficticias* muy utilizadas en la teoría de modelos y que suelen introducirse tradicionalmente para incorporar al modelo posibles variaciones estructurales ocurridas durante el período muestral, o efectos socioeconómicos que pudieran distinguir el comportamiento de unos individuos de otros.

En estos casos, se introduce en el modelo una variable discreta para cada una de las características que se pretenden tomar en consideración. A cada una de las posibles modalidades que puede presentar la característica se le asocia un valor numérico, y la *variable ficticia* así definida, se utiliza en la estimación del modelo como una variable explicativa más. Si se pretende recoger en el modelo la idea de un posible cambio estructural en el instante t_o , entonces la variable ficticia correspondiente tomará el valor 0 para $1 < t < t_o$, y el valor 1 para $t_o < t < T$. Si se pretendiese discriminar entre los gastos en educación de las familias de una muestra dependiendo de su pertenencia a un medio rural o urbano, podría definirse una variable ficticia que toma el valor 0 si la familia vive en un medio rural, y el valor 1 si vive en un medio urbano.

Hay otras variables *discretas* que podrían ser influyentes en el comportamiento de la variable endógena sin ser variables ficticias. Por ejemplo, el número de hijos es importante al analizar el gasto en educación de una determinada familia. Conviene recordar que variables continuas como la edad o los ingresos de un individuo, se transforman en ocasiones en variables discretas debido al diseño de la muestra usada en la recogida de datos.

Por otra parte, son muchos los problemas y cuestiones de interés en economía, demografía, sociología, epidemiología, biología, medicina y otras ciencias en los que es la variable endógena la que puede no ser continua. A veces, ni siquiera esta variable es cuantificable, como ocurre cuando se pretende cuantificar el tipo de transporte que cada persona de un determinado colectivo toma diariamente cuando acude a su trabajo. En tal caso pueden considerarse como alternativas la utilización de coche particular o de transporte público, pudiendo introducir la variable endógena que vale 0 o 1 según que el medio de transporte utilizado sea privado o público.

Otros ejemplos de variable endógena (dependiente) cualitativa podría ser un modelo que trate de explicar el nivel de estudios de los individuos en función de sus características personales, o un modelo que tratase de explicar el votar o no a un determinado partido político en elecciones según determinadas variables que influyen en ello, el tipo de escuela (privada o pública) a la que envían las familias a los hijos, en función de variables como los ingresos, la edad, etc., la devolución o no en fecha de vencimiento de un crédito concedido por una entidad bancaria en función de determinadas características de los deudores.

Los modelos de regresión más comunes exigen que la distribución de la variable dependiente sea normal. Cuando dicha variable no es normal pero sí continua, se pueden intentar, para poder aplicar dichos modelos, transformaciones de la misma que normalicen su función de probabilidad. Pero esta estrategia, sin embargo, no sirve para variables discretas. En este Capítulo se va a introducir una ampliación en la teoría de modelos de regresión consistente en estudiar un modelo de regresión aplicable a variables dependientes discretas binomiales.

Se dice que un proceso es binomial cuando sólo tiene dos posibles resultados: "éxito" y "fracaso", siendo la probabilidad de cada uno de ellos constante a lo largo de una serie de repeticiones. A la variable número de éxitos en n repeticiones se le denomina variable binomial. A la variable resultado de un sólo ensayo y, por tanto, con sólo dos valores: 0 para fracaso y 1 para éxito, se le denomina *binomial puntual* o de *Bernouilli*. La terminología de "éxito" y "fracaso" proviene de los juegos de azar (una apuesta en un juego de azar sólo tiene dos resultados, se gana o se pierde) y se mantiene, aunque no es muy afortunada en otros contextos. La presencia o ausencia de un cierto carácter biológico, por ejemplo, que una pareja tenga un descendiente con ojos negros, es un fenómeno binomial (sólo dos resultados y la misma probabilidad en sucesivas repeticiones) sin que haya ninguna connotación de éxito en dicha presencia, incluso en Ciencias de la Salud y para el proceso, también binomial, de padecer, o no, una cierta enfermedad parece claramente inapropiado denominar éxito a enfermar.

Recuérdese que la función densidad de probabilidad de una variable binomial Y de parámetros n y p es la siguiente:

$$f(y) = \binom{n}{y} p^y (1-p)^{1-y}$$

La función de densidad para una variable binomial puntual (o variable de Bernouilli) viene dada por:

$$f(y) = p^y(1-p)^{1-y}$$

Un proceso binomial puntual está, pues, caracterizado por la probabilidad de éxito, representada por p (es el único parámetro de su función de probabilidad), y por la probabilidad de fracaso que se representa por q . Evidentemente, ambas probabilidades están relacionadas por $p+q=1$. En ocasiones, se usa el cociente p/q , denominado "odds" (o "ventaja") que indica cuánto más o menos probable es el éxito que el fracaso. El *odds* se define como:

$$Odds = \text{Probabilidad de éxito} / \text{Probabilidad de fracaso} = p/q$$

Por ejemplo; si la probabilidad de éxito es 0,75, la de fracaso será 0,25, con lo que el *odds* valdrá $0,75/0,25=3$. Esto puede interpretarse diciendo que el éxito es tres veces más probable que el fracaso. Siempre que el *odds* sea mayor que 1 la probabilidad de éxito será mayor que la fracaso, si el *odds* es menor que el fracaso es más probable que el éxito y si el *odds* es la unidad éxito y fracaso son igualmente probables.

También es posible obtener la probabilidad de éxito conociendo el *odds*, ya que se tiene lo siguiente:

$$Odds = p/q = p/(1-p) \Rightarrow p = odds/(odds+1)$$

Por ejemplo; si el *odds* vale 3, la probabilidad de éxito es $p=3/4=0.75$.

Otro concepto, que se utiliza para comparar el *odds* de dos sucesos es el "odds ratio" (o "ventaja comparativa"), que se define como el cociente de los *odds* para los dos sucesos, y por lo tanto mide cuánto más ventajoso es el éxito sobre el fracaso en el primer suceso que en el segundo.

Suele denotarse el *odds* para un suceso por la letra griega π , mientras que el *odds ratio* de dos sucesos se denota por la letra griega θ . Si $\theta = \pi_1/\pi_2$ es la unidad, los *odds* para los dos sucesos son iguales, con lo que las probabilidades de éxito en los dos sucesos serán iguales, ya que $p=odds/(odds+1)$ vale lo mismo para los dos sucesos. Si $\theta = \pi_1/\pi_2 > 1$ el éxito es más ventajoso sobre el fracaso en el primer susceso. Si $\theta = \pi_1/\pi_2 < 1$ el éxito es más ventajoso sobre el fracaso en el segundo susceso. El *odds ratio* se denomina también "riesgo relativo".

EL MODELO DE REGRESIÓN LOGÍSTICA

Un modelo de regresión con variable dependiente binomial (modelo logístico o modelo de regresión logística) será un modelo que permita estudiar si dicha variable discreta depende o no, de otra u otras variables. Si una variable binomial de parámetro p es independiente de otra variable X , se cumple $(p|X=x) = p$, para cualquier valor x de la variable X .

Por consiguiente, un modelo de regresión con variable dependiente binomial y una única variable independiente X se materializa en una función en la que p aparece dependiendo de X y de unos coeficientes cuya investigación permite abordar la relación de dependencia. Para una única variable independiente X , el modelo de regresión logística toma la forma:

$$\ln(p/q | X = x) = \alpha_0 + \alpha_1 X, \text{ o de forma simplificada: } \ln(p/q) = \alpha_0 + \alpha_1 X$$

donde \ln significa logaritmo neperiano, α_0 y α_1 son constantes y X una variable que puede ser aleatoria o no, continua o discreta. Este modelo se puede fácilmente generalizar para k variables independientes, dando lugar al *modelo logístico múltiple*, que se expresa como sigue:

$$\ln(p/q) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k$$

Hemos definido el *modelo logístico* como el logaritmo del *odds* para el suceso que representa la variable aleatoria binomial puntual dependiente del modelo.

Hay varias razones para plantear el modelo con el logaritmo de *odds*, en lugar de plantearlo simplemente con la probabilidad de éxito o con el *odds*. En primer lugar, el campo de variación de $\ln(p/q)$ es todo el campo real (de $-\infty$ a ∞), mientras que, para p el campo es sólo de 0 a 1 y para p/q es de 0 a ∞ . Por lo tanto, con el modelo definido en función del logaritmo de *odds* no hay que poner restricciones a los coeficientes que complicarían su estimación. Por otro lado, y más importante, en los modelos en función del logaritmo de *odds* los coeficientes son fácilmente interpretables en términos de independencia o asociación entre las variables, como se verá más adelante.

El modelo logístico se puede escribir de otras formas equivalentes que para ciertas aplicaciones son más cómodas de manejar. Tenemos:

$$\begin{aligned} \ln(p/q) = \alpha_0 + \alpha_1 X &\Leftrightarrow \ln\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 X \Leftrightarrow \frac{p}{1-p} = e^{\alpha_0 + \alpha_1 X} \Leftrightarrow \\ p &= \frac{e^{\alpha_0 + \alpha_1 X}}{1 + e^{\alpha_0 + \alpha_1 X}} \Leftrightarrow p = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X)}} \end{aligned}$$

Estas dos últimas expresiones, si son conocidos los coeficientes α_0 y α_1 , permiten calcular directamente la probabilidad del proceso binomial para los distintos valores de la variable X .

A la función: $f(z) = \frac{1}{1+e^{-z}}$ se le denomina función logística. El modelo de regresión logística modeliza la probabilidad de un proceso binomial como la función logística de una combinación lineal de la variable dependiente.

El modelo de regresión logística múltiple tendrá la expresión:

$$p = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_k X_k)}}$$

INTERPRETACIÓN DE LOS COEFICIENTES DEL MODELO

Cuando $X=0$ el modelo logístico $\ln(p/q) = \alpha_0 + \alpha_1 X$ queda de la forma: $\ln(p/q | X=0) = \alpha_0$, con lo que ya podemos decir que **α_0 es el logaritmo de odds cuando la variable independiente es cero.**

Cuando $X=1$ y la variable X sólo puede tomar los valores 0 y 1 se tiene que:

$$\ln(p/q | X=1) = \alpha_0 + \alpha_1 = \ln(p/q | X=0) + \alpha_1 \Leftrightarrow$$

$$\alpha_1 = \ln(p/q | X=1) - \ln(p/q | X=0) = \ln \frac{p/q | X=1}{p/q | X=0}$$

Por lo tanto, se puede decir que **α_1 es el logaritmo del cociente de los odds para los valores de la variable X, supuesto que X sólo pueda tomar los valores 0 y 1.** Por lo tanto, **α_1 es el odds ratio para los dos únicos valores 0 y 1 de la variable X.**

Además, si la variable binomial es independiente la variable X, ambos *odds* son iguales, por lo tanto el *odds ratio* es 1 y su logaritmo es cero. Por lo tanto, **para estudiar con un modelo logístico la independencia de las variables basta comprobar si el coeficiente α_1 es cero.**

Decir que α_0 es el logaritmo de *odds* cuando la variable independiente es cero es lo mismo que decir que e^{α_0} es el *odds* cuando $X=0$.

Decir que α_1 es el logaritmo del cociente de los *odds* para los dos valores de la variable X es lo mismo que decir e^{α_1} es el *odds ratio* entre $X=1$ y $X=0$.

Si la variable X puede tomar más de dos valores, evidentemente se sigue manteniendo que e^{α_0} es el odds cuando $X=0$. Además, en este caso, e^{α_1} es el *odds ratio* para el aumento en una unidad en la variable X (lo que implica que el *odds ratio* para el aumento en una unidad en la variable X es constante en el modelo de regresión logística).

Cualesquiera que sean los valores que tome la variable X se cumple que: $\ln(p/q | X = x_0) = \alpha_0 + \alpha_1 x_0$ y $\ln(p/q | X = x_1) = \alpha_0 + \alpha_1 x_1$, con lo que:

$$\underbrace{\ln(p/q | X = x_1) - \ln(p/q | X = x_0)}_{=} = (\alpha_0 + \alpha_1 x_1) - (\alpha_0 + \alpha_1 x_0) = \alpha_1(x_1 - x_0) = \alpha_1 \delta$$

$$\ln \frac{p/q | X = x_1}{p/q | X = x_0} = \alpha_1 \delta \Rightarrow \text{odds ratio} = e^{\alpha_1 \delta} = (e^{\alpha_1})^\delta$$

Por lo tanto, para un aumento de la variable X desde x_0 a x_1 (es decir, para un aumento en X de valor $\delta=x_1-x_0$) se tiene que $e^{\alpha_1 \delta}$ es el *odds ratio* para el aumento en δ unidades en la variable X. Cuando $\delta=1$, se tiene que el *odds ratio* para el aumento en una unidad en la variable X es constante e igual a e^{α_1} en el modelo de regresión logística, es decir, que **el logaritmo del odds ratio para el aumento en una unidad en la variable X es constante y precisamente igual al coeficiente α_1 del modelo.**

ESTIMACIÓN DE LOS COEFICIENTES

El método de los mínimos cuadrados, clásico en la estimación de los coeficientes de los modelos de regresión no es aplicable al modelo logístico, ya que dicho método se basa en la normalidad de la variable dependiente, que en este caso no se cumple. Por otra parte, cuando $q=0$, es imposible calcular $\ln(p/q)$. Se tratará entonces de utilizar el método de máxima verosimilitud.

Comenzaremos considerando el caso más simple con una sola variable independiente X. Tomamos una muestra de n observaciones (y_i, x_i) para la variable puntual binomial dependiente Y y para la variable independiente X. La variable Y toma valores y_i que sólo pueden ser 1 con probabilidad p_i o 0 con probabilidad $1-p_i$. Como x_i depende de p_i a través del modelo logístico tenemos:

$$p_i = \frac{e^{\alpha_0 + \alpha_1 X_i}}{1 + e^{\alpha_0 + \alpha_1 X_i}}$$

La función de verosimilitud para una variable binomial puntual es:

$$L(p_i | y_i) = (p_i)^{y_i} (1-p_i)^{1-y_i}$$

y para n observaciones independientes la función de verosimilitud de la muestra será:

$$L(p_1, \dots, p_n | y_1, \dots, y_n) = \prod_{i=1}^n (p_i)^{y_i} (1-p_i)^{1-y_i}$$

y al representar p_i por el modelo logístico tendremos ya la expresión de la función de verosimilitud para la muestra como función de los parámetros a estimar:

$$\begin{aligned} L(p_1, \dots, p_n | y_1, \dots, y_n) &= \prod_{i=1}^n \left(\frac{e^{\alpha_0 + \alpha_1 X_i}}{1 + e^{\alpha_0 + \alpha_1 X_i}} \right)^{y_i} \left(1 - \frac{e^{\alpha_0 + \alpha_1 X_i}}{1 + e^{\alpha_0 + \alpha_1 X_i}} \right)^{1-y_i} = \\ &= \prod_{i=1}^n \left(\frac{e^{\alpha_0 + \alpha_1 x_i}}{1 + e^{\alpha_0 + \alpha_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\alpha_0 + \alpha_1 x_i}} \right)^{1-y_i} = \frac{e^{\alpha_0 \sum_{i=1}^n y_i + \alpha_1 \sum_{i=1}^n x_i y_i}}{\prod_{i=1}^n 1 + e^{\alpha_0 + \alpha_1 x_i}} = L(\alpha_0, \alpha_1) \end{aligned}$$

y como suele ser usual en máxima verosimilitud, maximizaremos el logaritmo de la función $L(\alpha_0, \alpha_1)$ en vez de la función misma. Los parámetros estimados del modelo serán los valores de α_0 y α_1 , que maximicen la función $\ln L(\alpha_0, \alpha_1)$.

ESTIMACIÓN POR INTERVALOS Y CONTRASTES DE HIPÓTESIS SOBRE LOS COEFICIENTES

Según el teorema central del límite, los estimadores por máxima verosimilitud de los parámetros del modelo logístico son asintóticamente normales y su matriz de varianzas covarianzas es perfectamente calculable a partir del algoritmo de maximización de la función de verosimilitud (método de Newton Rhampson).

De esta forma, un intervalo de confianza al $(1-\alpha)\%$ para el estimador del coeficiente α_i del modelo será:

$$\hat{\alpha}_i \pm Z_{\alpha/2} \hat{\sigma}(\hat{\alpha}_i)$$

Hay que tener presente que los estimadores habituales que miden asociación entre variables son los *odds ratio*, por lo tanto interesa dar los intervalos de confianza para los *odds ratio*, que evidentemente serán:

$$e^{\hat{\alpha}_i \pm Z_{\alpha/2} \hat{\sigma}(\hat{\alpha}_i)}$$

El estadístico para el contraste:

$$H_0: \alpha_i = a$$

$$H_1: \alpha_i \neq a$$

Será: $Z = \frac{\hat{\alpha}_i - \alpha}{\hat{\sigma}(\hat{\alpha}_i)} \rightarrow N(\alpha, \hat{\sigma}(\hat{\alpha}_i))$ y región crítica $|Z| > Z_\alpha/2$

También suele utilizarse para el contraste el estadístico de Wald definido como $W = Z^2$ y cuya distribución es una chi-cuadrado con 1 grado de libertad. La región crítica de este contraste es $W > \chi^2_\alpha$.

En el modelo logístico es muy interesante contrastar la hipótesis $\alpha_i = 0$ para $i=1,\dots,k$, porque no rechazar esta hipótesis para un valor de i implica que la variable Y no depende X_i , y por lo tanto esta última no debería figurar en el modelo.

También suele utilizarse el contraste de la razón de verosimilitudes, basado en el estadístico $-2\text{Log}(L_0/L_1)$ donde L_0 es el máximo de la función de verosimilitud bajo la hipótesis nula y L_1 es el máximo de la función de verosimilitud bajo la hipótesis alternativa. Este estadístico tiene una distribución chi-cuadrado con grados de libertad igual al número de parámetros bajo la hipótesis nula.

MODELOS PROBIT Y LOGIT

Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que cumple esta condición se denomina **modelo lineal de probabilidad**, y tiene la forma funcional:

$$P_i = F(x_i, \beta) + u_i$$

Se observa que si F es la función de distribución de una variable aleatoria entonces P varía entre cero y uno.

En el caso particular en que la función F es la función logística estaremos ante el **modelo Logit**, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} + u_i$$

En el caso particular en que la función F es la función de distribución de una normal unitaria estaremos ante el **modelo Probit**, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x_i \beta} e^{-\frac{t^2}{2}} dt + u_i$$

ANÁLISIS DE LA SUPERVIVENCIA

En los últimos años se han popularizado los modelos de probabilidad que intentan analizar, bien las series de tiempos de fallo en un proceso industrial, bien las series de tiempos de supervivencia de un grupo de individuos, bien las series de duraciones de ciertos procesos, bien las series de tiempos que se permanece empleado en la misma empresa, de tiempos de permanencia en la universidad, de tiempos desde el matrimonio a la llegada del primer hijo, etc. Tiene cierto interés conocer los años de supervivencia tras una intervención quirúrgica (un trasplante de corazón, por ejemplo) o tras el diagnóstico de una enfermedad (el SIDA u otras). Quizás se podría pretender determinar el tiempo que transcurre desde la administración de determinado fármaco y la desaparición de sus efectos. Obsérvese que todos los ejemplos se refieren a un suceso único, irrepetible sobre los individuos y no recurrente. El análisis de este tipo de series, que normalmente siempre son series de tiempo, suele conocerse con el nombre genérico de **análisis de la supervivencia** (bien sea referida a individuos, máquinas, etc.).

La variable de interés en el análisis de supervivencia es la longitud del período de tiempo que transcurre desde el principio de algún acontecimiento hasta el final del mismo, o hasta el momento en que ese acontecimiento es observado, lo que puede ocurrir antes de que el acontecimiento acabe. Los datos habitualmente se presentan como un conjunto de duraciones o supervivencias, t_1, t_2, \dots, t_n que no necesariamente tienen porqué empezar en el mismo punto del tiempo.

Una característica inherente al análisis de supervivencia es la **censura**. Se dice que **los datos están censurados** si no se pueden observar por completo. Consíderese por ejemplo el análisis del tiempo que transcurre entre el diagnóstico de un determinado tipo de cáncer en un grupo de pacientes y la muerte de los mismos. Los pacientes son observados cada seis meses, empezando justo en el momento en que se les diagnosticó el cáncer. Por el momento supóngase que a todos los pacientes se les diagnosticó el cáncer el mismo día. Tras seis meses algunos pacientes han muerto y otros no. Para los pacientes sobrevivientes la duración, o supervivencia, es por lo menos igual al período observado, $t_i = 6$ meses, pero no es igual a él. Este tipo de censura, la más habitual, se conoce como **censura por la derecha** (tiempo de supervivencia real mayor que el observado).

Es posible, así mismo, que exista **censura por la izquierda**, en cuyo caso el tiempo de supervivencia real es menor que el observado. Supongamos por ejemplo que estamos interesados en la supervivencia de un grupo de pacientes con síntomas de un determinado tipo de cáncer, hayan sido o no diagnosticados. En este caso algunos pacientes pueden haber muerto antes de que se les diagnosticase. Tales pacientes presentan censura por la izquierda.

La censura también puede ser **censura de intervalo**, por cuanto se conoce que el evento irrepetible ha ocurrido en un intervalo de tiempo determinado. Supongamos ahora que algunos de los pacientes sobrevivientes seis meses después de serles diagnosticado el cáncer, han muerto en la observación, un año después. Existe entonces una censura de intervalo, entre seis meses y año.

TABLAS DE MORTALIDAD

En el análisis de la supervivencia lo primero que se analiza es la tabla de mortalidad (o tabla de vida). Esta tabla contiene distintas informaciones descriptivas sobre la evolución de las observaciones, entre las que tenemos las siguientes:

- **Intervalos (Intervals):** Aparecen los límites inferior y superior de los intervalos de tiempo en que se ha dividido la serie, y en cada uno de los cuales se han registrado el correspondiente número de fallos (fallecimientos, averías, etc.).
- **Número de fallos en cada intervalo (Number of Failures).**
- **Número de abandonos (Number Withdrawn):** Individuos que no han llegado al final de periodo debido a abandono (mortalidad experimental) y no a causa de producirse el evento terminal (fallecimiento). Se puede observar cómo puede haber observaciones que no han finalizado el primer intervalo y sin embargo no han fallecido. También hay individuos sobre los que no se tiene información completa para ciertos intervalos de tiempo. Por tanto, no se pueden contabilizar como fallecimientos, ni como vivos, sino como abandonos.
- **Número de observaciones expuestas a riesgo (Number at Risk).**
- **Funciones de supervivencia:** En las tablas de vida se incluyen funciones de supervivencia (*survival functions*) con objeto de facilitar la interpretación de los resultados. Las principales funciones de supervivencia son: la **función de supervivencia acumulada** estimada, la **función de riesgo** (o azar) estimada y la **función de densidad** estimada para la distribución de los tiempos de la serie de vida en estudio.

Más formalmente, sea T una variable aleatoria continua no negativa con función de densidad $f(t)$, que representa el tiempo de supervivencia (por ejemplo, de un paciente, de una máquina, etc.). Su función de distribución, o función de probabilidad acumulada es $F(t) = \text{Prob}(T \leq t)$. La función de supervivencia $S(t)$ se define como la probabilidad de supervivencia hasta t , o sea $S(t) = \text{Prob}(T \geq t) = 1 - F(t)$. La función de riesgo o tasa de azar $h(t)$ se define como la probabilidad de que un individuo, que se supone vivo en el instante t , sobreviva al siguiente intervalo de tiempo lo suficientemente pequeño, o sea, $h(t)$ es la función de densidad condicional en t dada la supervivencia hasta t y se tiene $h(t) = f(t)/S(t)$.

La función de densidad, la función de riesgo y la función de supervivencia están relacionadas mediante $f(t) = S(t)h(t)$ y $h(t) = -d\ln S(t)/dt$. Otra función de interés es la función integrada o acumulada de riesgo $H(t) = -\ln S(t)$. En ocasiones se dice también que la variable T representa el tiempo de fallo, sobre todo en teoría de la fiabilidad.

El objetivo del análisis de supervivencia es estimar las funciones de supervivencia y de riesgo a partir de los tiempos de supervivencia observados. Existen dos métodos principales para el cálculo de estas funciones: a) método actuarial de Berkson y Gage (1950) y b) método del producto de Kaplan y Meyer (1958). Una síntesis de estos métodos puede verse en Pardell, Cobo y Canela (1986).

Estimaciones no paramétricas de la función de supervivencia

Este procedimiento más común que permite realizar estimaciones no paramétricas de la función de supervivencia, basadas en la función de supervivencia empírica de la muestra, es el **método del límite producto de Kaplan Meier** (la función de supervivencia empírica es $S_m(t) = \text{Número de individuos con tiempo de supervivencia mayor o igual que } t \text{ dividido entre el número total de individuos}$, y se tiene que $S_m(t) = 1 - F_m(t)$ donde $F_m(t)$ es la función de distribución empírica). Se usa para obtener probabilidades de supervivencia para datos multicensados y también se usa en ensayos clínicos para estudiar y comparar tasas de supervivencia de pacientes bajo diferentes tratamientos.

Estimaciones paramétricas de la función de supervivencia

Las aproximaciones no paramétricas no necesitaban especificar tipo de distribución de probabilidad para los tiempos de supervivencia. De este modo, la función de riesgo tampoco necesita ser especificada permitiendo, por tanto, una gran flexibilidad en el análisis.

Ahora bien, cuando los datos respondan efectivamente a una determinada distribución de probabilidad, las inferencias basadas en la parametrización de dicha distribución serán más precisas o eficientes. Si la distribución de probabilidad asumida es correcta, los errores estándar de los estimadores en las aproximaciones paramétricas son menores. Además estas aproximaciones permiten realizar inferencias poblacionales no limitándose a la muestra analizada como en el caso de las alternativas puramente no paramétricas.

Supongamos ahora que los datos siguen un modelo de probabilidad determinado. El modelo más sencillo es el que supone que la tasa de riesgo no varía en el tiempo, es decir $h(t)$ es constante. En este caso, la probabilidad condicionada a estar vivo en t de que un individuo muera en un intervalo de tiempo determinado (lo suficientemente pequeño) será la misma con independencia del momento en el que se observe el individuo.

Esta característica se conoce como **pérdida de memoria**. La función de riesgo puede representarse por $h(t) = \lambda$ para $0 \leq t \leq \infty$ siendo λ una constante positiva. Como $-d\ln S(t)/dt = h(t) = \lambda \Rightarrow S(t) = K e^{-\lambda t}$, y como $S(0) = 1$ entonces $K = 1$, con lo que $S(t) = e^{-\lambda t}$ y estamos ante la **distribución exponencial** para los datos, porque la función de densidad es $f(t) = h(t)S(t) = \lambda e^{-\lambda t}$.

El problema de la distribución exponencial es que, salvo en procesos industriales, es difícilmente sostenible que la supervivencia se defina por una tasa de riesgo constante. Por este motivo se han propuesto otras distribuciones alternativas, entre las cuales, la más utilizada es la **distribución de Weibull**, que supone que la tasa de riesgo toma la forma $h(t) = \lambda\gamma t^{\gamma-1}$ para $0 \leq t < \infty$ y donde los parámetros λ (**parámetro de escala**) y γ (**parámetro de forma**) son constantes positivas. Si $\gamma = 1$, la función de riesgo es constante, con lo que los tiempos de supervivencia siguen una distribución exponencial. Para otros valores de γ , la función de riesgo crece o decrece de forma monótona (no cambia de dirección). Para el valor ya conocido de $h(t)$ tenemos:

$$S(t) = e^{-\lambda t^\gamma} \Rightarrow f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^\gamma} \quad (\text{función de densidad de la variable Weibull}).$$

Existe otros modelos típicos en el análisis de la supervivencia, como por ejemplo, el **modelo log-logístico**, cuya función de riesgo es $h(t) = (e^{\theta_1 t^{\kappa-1}})/(1 + e^{\theta_1 t^{\kappa}})$, siendo la función de supervivencia $S(t) = (1 + e^{\theta_1 t^{\kappa}})^{-1}$ y siendo la función de densidad $f(t) = (e^{\theta_1 t^{\kappa-1}})/(1 + e^{\theta_1 t^{\kappa}})^2$, que es la función de densidad de una variable log-logística.

REGRESIÓN DE COX

Del mismo modo que las tablas de mortalidad y el análisis de supervivencia de Kaplan-Meier, la regresión de Cox es un método para crear modelos para datos de tiempos de espera hasta un evento con casos censurados presentes. Sin embargo, la regresión de Cox permite incluir en los modelos variables predictoras (covariables). La regresión de Cox gestionará los casos censurados correctamente y proporcionará las estimaciones de los coeficientes para cada una de las covariables, permitiendo evaluar el impacto de múltiples covariables en el mismo modelo. Además, es posible utilizar la regresión de Cox para examinar el efecto de covariables continuas.

La función de azar para supervivencia en el tiempo de los miembros de una población viene dada según Cox por:

$$H(t|x) = h_0(t)\exp(\beta' x(t))$$

donde $x(t)$ es un vector de covariables posiblemente dependientes del tiempo y β es un vector de parámetros de regresión a estimar. La función de supervivencia será:

$$S(t|x) = S_0(t)\exp(\beta' x(t)) \quad \text{dónde } S_0(t) = \int h_0(u)du$$

La regresión de Cox permite contrastes de hipótesis lineales a cerca de los parámetros de la regresión, cálculo de intervalos de confianza y análisis de los residuos.

SPSS Y LOS MODELOS DE ELECCIÓN DISCRETA LOGIT Y PROBIT. REGRESIÓN DE COX

SPSS Y LA REGRESIÓN LOGÍSTICA

SPSS incorpora un procedimiento que implementa el análisis de regresión logística. La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de variables predictoras. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de las ventajas (*odds ratio*) de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante.

Como ejemplo podemos preguntarnos: ¿Qué características del organismo humano son factores efectivamente interviniéntes en el crecimiento? Dada una muestra de niños y niñas a los que se mide la edad, sexo y la distancia del centro de la pituitaria a la fisura ptérigo-maxilar, se puede construir un modelo para predecir la presencia o ausencia de crecimiento según el sexo en la muestra de personas. El modelo puede utilizarse posteriormente para derivar estimaciones de la razón de las ventajas para cada uno de los factores y así indicarle, por ejemplo, cuánto más probable es que las características de crecimiento indicadas intervengan más en los niños que en las niñas.

También podríamos utilizar un modelo de regresión logística para estudiar qué factores significan riesgo de enfermedad cardiovascular. Dada una muestra de pacientes a los que se mide la situación de fumador, dieta, ejercicio, consumo de alcohol, y estado de enfermedad cardiovascular, se puede construir un modelo utilizando las cuatro variables de estilo de vida para predecir la presencia o ausencia de enfermedad cardiovascular en una muestra de pacientes.

El modelo puede utilizarse posteriormente para derivar estimaciones de la razón de las ventajas para cada uno de los factores y así indicarle, por ejemplo, cuánto más probable es que los fumadores desarrollen una enfermedad cardiovascular frente a los no fumadores.

Para cada análisis se obtienen estadísticos como los *Casos totales*, *Casos seleccionados* y *Casos válidos*. Para cada variable categórica se obtiene la codificación de los parámetros. Para cada paso se obtienen las variables introducidas o eliminadas, historial de iteraciones, $-2 \log$ de la verosimilitud, bondad de ajuste, estadístico de bondad de ajuste de Hosmer-Lemeshow, chi-cuadrado del modelo, chi-cuadrado de la mejora, tabla de clasificación, correlaciones entre las variables, gráfico de las probabilidades pronosticadas y los grupos observados y chi-cuadrado residual. Para cada variable de la ecuación se obtiene: coeficiente (B), error típico de B, estadístico de Wald, R, razón de las ventajas estimada ($\exp(B)$), intervalo de confianza para $\exp(B)$, log de la verosimilitud si el término se ha eliminado del modelo. Para cada variable que no esté en la ecuación se obtiene: estadístico de puntuación y R. Para cada caso se obtiene: grupo observado, probabilidad pronosticada, grupo pronosticado, residuo y residuo tipificado. Puede estimar modelos utilizando la entrada en bloque de las variables o cualquiera de los siguientes métodos por pasos: condicional hacia adelante, LR hacia adelante, Wald hacia adelante, condicional hacia atrás, LR hacia atrás o Wald hacia atrás.

Para realizar un análisis de regresión logística binaria, elija en los menús *Analizar* → *Regresión* → *Logística binaria* (Figura 20-1) y seleccione las variables y las especificaciones para el análisis (Figura 20-2). Previamente es necesario cargar en memoria el fichero de nombre CRECIMIENTO mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene datos sobre características del organismo humano que son factores efectivamente intervinientes en el crecimiento de los niños. La variable cualitativa dependiente va a ser *género* (niño o niña) y las variables independientes son la distancia del centro de la pituitaria a la fisura ptérigo-maxilar (*distanci*) y la edad (*edad*). Se ajustará un modelo que prediga el género según los valores de las variables independientes.

En cuanto a los datos, la variable dependiente debe ser dicotómica. Las variables independientes pueden estar a nivel de intervalo o ser categóricas. Si son categóricas, deben ser variables *dummy* o estar codificadas como indicadores (existe una opción en el procedimiento para recodificar automáticamente las variables categóricas).

En cuanto a supuestos, la regresión logística no se basa en supuestos distribucionales en el mismo sentido en que lo hace el análisis discriminante. Sin embargo, la solución puede ser más estable si los predictores tienen una distribución normal multivariante. Adicionalmente, al igual que con otras formas de regresión, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y a errores típicos inflados. El procedimiento es más eficaz cuando la pertenencia a grupos es una variable categórica auténtica; si la pertenencia al grupo se basa en valores de una variable continua (por ejemplo "CI alto" en contraposición a "CI bajo"), deberá considerar el utilizar la regresión lineal para aprovechar la información mucho más rica ofrecida por la propia variable continua.

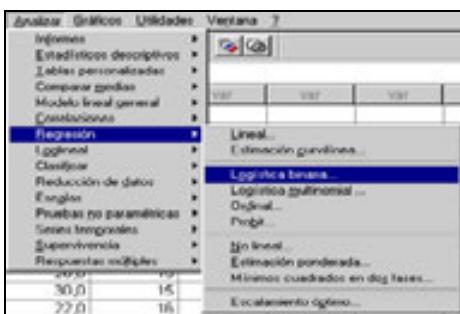


Figura 20-1



Figura 20-2

En los campos *Dependiente* y *Covariables* de la Figura 20-2 se introducen las variables dependiente e independientes del modelo. En el botón *Categórica* (Figura 20-3) puede especificar los detalles sobre cómo el procedimiento *Regresión logística* manipulará las variables categóricas. El campo *Covariables* contiene una lista de todas las covariables especificadas en el cuadro de diálogo principal para cualquier capa, bien por ellas mismas o como parte de una interacción. Si alguna de éstas son variables de cadena o son categóricas, sólo puede utilizarlas como covariables categóricas. En el campo *Covariables categóricas* se introduce la lista de las variables identificadas como categóricas. Cada variable incluye una notación entre paréntesis indicando el esquema de codificación de contraste que va a utilizarse. Las variables de cadena (señaladas con el símbolo < a continuación del nombre) estarán presentes ya en la lista *Covariables categóricas* por defecto. Seleccione cualquier otra covariable categórica de la lista *Covariables* y muévala a la lista *Covariables categóricas*. El botón *Cambiar el contraste* le permite cambiar el método de contraste.

Los métodos de contraste disponibles son: *Desviación* (cada categoría de la variable predictora, excepto la categoría de referencia, se compara con el efecto global), *Simple* (cada categoría de la variable predictora, excepto la misma categoría de referencia, se compara con la categoría de referencia), *Diferencia* o *contrastos de Helmert inversos* (cada categoría de la variable predictora, excepto la primera categoría, se compara con el efecto promedio de las categorías anteriores), *Helmert* (cada categoría de la variable predictora, excepto la última categoría, se compara con el efecto promedio de las categorías subsiguientes), *Repetida* (cada categoría de la variable predictora, excepto la primera categoría, se compara con la categoría que la precede), *Polinómico* (contrastos polinómicos ortogonales en los que se supone que las categorías están espaciadas equidistantemente y sólo están disponibles para variables numéricas) e *Indicador* (los contrastes indican la presencia o ausencia de la pertenencia a una categoría y la categoría de referencia se representa en la matriz de contraste como una fila de ceros). Si selecciona *Desviación*, *Simple* o *Indicador*, elija *Primera* o *Última* como categoría de referencia. Observe que el método no cambia realmente hasta que se pulsa en *Cambiar*. Las covariables de cadena deben ser covariables categóricas. Para eliminar una variable de cadena de la lista *Covariables categóricas*, debe eliminar de la lista *Covariables* del cuadro de diálogo principal todos los términos que contengan la variable.

En el botón *Opciones* (Figura 20-4) puede especificar varias opciones para el análisis de regresión logística. La opción *Estadísticos y gráficos* le permite solicitar estadísticos y gráficos. Las opciones disponibles son *Gráficos de clasificación*, *Bondad de ajuste de Hosmer-Lemeshow*, *Listado de residuos por caso*, *Correlaciones de estimaciones*, *Historial de iteraciones* e *IC para exp(B)*. Seleccione una de las alternativas del grupo *Mostrar* para mostrar los estadísticos y los gráficos *En cada paso* o bien sólo para el modelo final, *En el último paso*. La opción *Probabilidad para el método por pasos* le permite controlar los criterios por los cuales las variables se introducen y se eliminan de la ecuación. Puede especificar criterios para la entrada o para la salida de variables. La opción *Punto de corte para la clasificación* le permite determinar el punto de corte para la clasificación de los casos. Los casos con valores pronosticados que han sobrepasado el punto de corte para la clasificación se clasifican como positivos, mientras que aquéllos con valores pronosticados menores que el punto de corte se clasifican como negativos. Para cambiar los valores por defecto, introduzca un valor comprendido entre 0,01 y 0,99. La opción *Nº máximo de iteraciones* le permite cambiar el número máximo de veces que el modelo itera antes de finalizar. La opción *Incluir constante en el modelo* le permite indicar si el modelo debe incluir un término constante. Si se desactiva, el término constante será igual a 0.

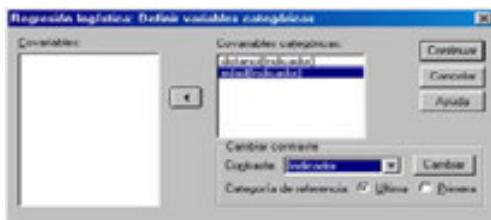


Figura 20-3



Figura 20-4

La opción *Seleccionar* (Figura 20-5) permite limitar el análisis a un subconjunto de casos que tengan un valor particular en una variable. Después de seleccionar esta opción elija una *Variable de selección* e introduzca un valor en *Establecer valor* para la variable de selección de casos. Los casos definidos por la regla de selección se incluyen en la estimación del modelo. Por ejemplo, si ha seleccionado una variable y la opción igual que y ha especificado 5 como valor, sólo se incluirán en el análisis aquellos casos para los cuales la variable seleccionada tenga un valor igual a 5. Los resultados de la clasificación y los estadísticos se generan tanto para los casos seleccionados como para los no seleccionados. De esta manera, se ofrece un mecanismo para clasificar los nuevos casos basándose en datos ya existentes; o también para realizar la partición de los datos en dos subconjuntos, uno de entrenamiento y otro de prueba, que permiten la validación del modelo generado.

La opción *Guardar* (Figura 20-6) permite guardar los resultados de la regresión logística como nuevas variables en el archivo de datos de trabajo. El campo *Valores pronosticados* guarda los valores pronosticados por el modelo. Las opciones disponibles son *Probabilidades* y *Grupo de pertenencia*. El campo *Influencia* guarda los valores de estadísticos que miden la influencia de los casos sobre los valores pronosticados.

Las opciones disponibles son: *De Cook*, *Valores de influencia* y *DfBeta(s)*. El campo *Residuos* guarda los residuos. Las opciones disponibles son: *No tipificados*, *Logit*, *Método de Student*, *Tipificados* y *Desviación*.

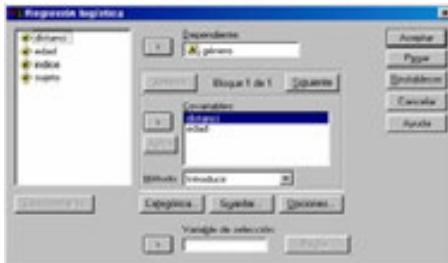


Figura 20-5

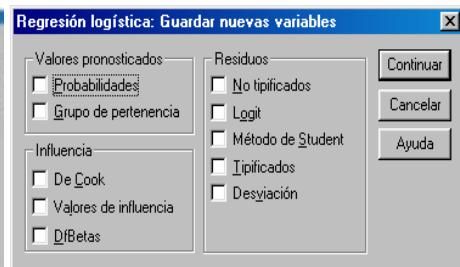


Figura 20-6

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 20-2 para obtener los resultados del análisis según se muestra en la Figura 20-7. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla.

En las Figuras 20-7 a 20-14 se presentan varias salidas tabulares y gráficas de entre las múltiples que ofrece el procedimiento.

LOGISTIC REGRESSION VAR=genero /METHOD=ENTER distancia edad /CLSPLOT /CASEWISE OUTLIER(2) /PRINT=GOODFIT CORR ITER(1) CI(95) /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .																				
Resumen del procesamiento de los casos																				
<table border="1"> <thead> <tr><th>Casos no ponderados^a</th><th>N</th><th>Porcentaje</th></tr> </thead> <tbody> <tr><td>Casos seleccionados</td><td>Incluidos en el análisis</td><td>108 100,0</td></tr> <tr><td></td><td>Casos perdidos</td><td>0 ,0</td></tr> <tr><td></td><td>Total</td><td>108 100,0</td></tr> <tr><td>Casos no seleccionados</td><td></td><td>0 ,0</td></tr> <tr><td>Total</td><td></td><td>108 100,0</td></tr> </tbody> </table> <p>a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.</p>			Casos no ponderados ^a	N	Porcentaje	Casos seleccionados	Incluidos en el análisis	108 100,0		Casos perdidos	0 ,0		Total	108 100,0	Casos no seleccionados		0 ,0	Total		108 100,0
Casos no ponderados ^a	N	Porcentaje																		
Casos seleccionados	Incluidos en el análisis	108 100,0																		
	Casos perdidos	0 ,0																		
	Total	108 100,0																		
Casos no seleccionados		0 ,0																		
Total		108 100,0																		
Codificación de la variable dependiente																				
<table border="1"> <thead> <tr><th>Valor original</th><th>Valor ínterno</th></tr> </thead> <tbody> <tr><td>Niña</td><td>0</td></tr> <tr><td>Niño</td><td>1</td></tr> </tbody> </table>			Valor original	Valor ínterno	Niña	0	Niño	1												
Valor original	Valor ínterno																			
Niña	0																			
Niño	1																			

Figura 20-7

Historial de iteraciones^{a,b,c}				
	-2 log de la verosimilitud	Coefficientes		
Iteración		Constante		
Paso 1	145,995	,370		
0 2	145,995	,375		
a. En el modelo se incluye una constante. b. -2 log de la verosimilitud inicial: 145,995 c. La estimación ha finalizado en el número de iteración 2 porque el logaritmo de la verosimilitud ha disminuido en menos de un,010 por ciento.				
Tabla de clasificación^{a,b}				
	Pronosticado			
	Género		Porcentaje correcto	
Observado	Género	Niña	Niño	
	Paso 0	0	44	,0
	Niño	0	64	100,0
	Porcentaje global			59,3
a. En el modelo se incluye una constante. b. El valor de corte es ,500				

Figura 20-8

Variables en la ecuación						
	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0 Constante	,375	,196	3,661	1	,056	1,455
Variables que no están en la ecuación						
Paso 0 Variables	DISTANCI	Puntuación	gl	Sig.		
	EDAD	,000	1	1,000		
Estadísticos globales		22,233	2	,000		

Figura 20-9

Bloque 1: Método = Introducir						
Historial de iteraciones ^{a,b,c,d}						
Iteración	-2 log de la verosimilitud	Coeficientes				
		Constante	DISTANCI	EDAD		
Paso 0	121,978	,576	,355	,234		
1	120,300	,7,559	,474	,306		
2	120,259	,7,943	,497	,319		
3	120,259	,7,955	,498	,319		
4	120,259	,7,955	,498	,319		
a. Método: Introducir						
b. En el modelo se incluye una constante.						
c. -2 log de la verosimilitud inicial: 145,995						
d. La estimación ha finalizado en el número de iteración 4 porque el logaritmo de la verosimilitud ha disminuido en menos de un ,010 por ciento.						
Pruebas omnibus sobre los coeficientes del modelo						
Paso	Chi-cuadrado	gl	Sig.			
Paso 1	25,736	2	,000			
Bloque	25,736	2	,000			
Modelo	25,736	2	,000			

Figura 20-10

Resumen de los modelos						
Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke			
1	120,259	,212	,286			
Prueba de Hosmer y Lemeshow						
Paso	Chi-cuadrado	gl	Sig.			
1	2,545	8	,960			

Tabla de contingencias para la prueba de Hosmer y Lemeshow

	Género = Niña		Género = Niño		Total
	Observado	Esperado	Observado	Esperado	
Paso 1	10	9,131	1	1,869	11
1	2	8	7,215	3	11
	3	6	6,694	6	12
	4	5	6,110	8	13
	5	4	4,104	6	10
	6	3	3,832	8	11
	7	3	3,031	8	11
	8	3	2,447	9	12
	9	2	1,205	9	11
	10	0	.231	6	6

Figura 20-11

Tabla de clasificación ^a						
Observado	Pronosticado					
	Género		Porcentaje correcto			
	Niña	Niño				
Paso 1	24	20	54,5			
Género	Niña					
Niño	10	54	84,4			
Porcentaje global			72,2			

a. El valor de corte es ,500

Variables en la ecuación						
	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 DISTANCI	,498	,120	17,129	1	,000	1,646
EDAD	-,319	,125	6,505	1	,011	,727
Constante	-,7,955	2,299	11,967	1	,001	,000
						I.C. 95,0% para EXP(B)
						Inferior Superior

a. Variable(s) introducida(s) en el paso 1: DISTANCI, EDAD.

Matriz de correlaciones			
	Constante	DISTANCI	EDAD
Paso 1 Constante	1,000	-,874	,148
DISTANCI	-,874	1,000	-,803
EDAD	,148	-,603	1,000

Figura 20-12

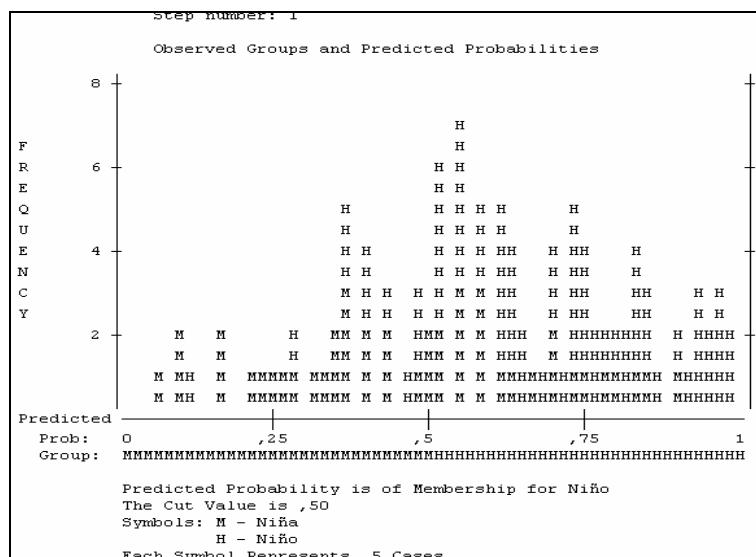


Figura 20-13

Listado por casos

Caso	Estado de selección ^a	Observado	Pronosticado	Grupo pronosticado	Variable temporal	
		Género			Resid	ZResid
43	S	M**	,897	H	-,897	-2,945
93	S	H**	,115	M	,885	2,775

a. S = Seleccionados, N = Casos no seleccionados y ** = Casos mal clasificados

b. Se listan los casos con residuos estudentizados mayores que 2,000

Figura 20-14

SPSS Y LA REGRESIÓN LOGÍSTICA MULTINOMIAL

SPSS incorpora un procedimiento que implementa el análisis de regresión logística multinomial. La opción Regresión logística multinomial resulta útil en aquellas situaciones en las que desee poder clasificar a los sujetos según los valores de un conjunto de variables predictoras. Este tipo de regresión es similar a la regresión logística, pero más general, ya que la variable dependiente no está restringida a dos categorías.

Como ejemplo podemos preguntarnos ¿Qué características del organismo humano son factores efectivamente interviniéntes en el crecimiento? Dada una muestra de niños y niñas a los que se mide la edad, sexo, la distancia del centro de la pituitaria a la fisura ptérico-maxilar y el índice de crecimiento, se puede construir un modelo para predecir el crecimiento según el índice (variable multinomial) en la muestra de personas. El modelo puede utilizarse posteriormente para derivar estimaciones de la razón de las ventajas para cada uno de los factores y así indicarle, por ejemplo, cuánto más probable es que las características de crecimiento indicadas intervengan más en los niños que presentan un índice u otro.

Otro ejemplo de aplicación se define a continuación: para conseguir una producción y distribución de películas más eficaz, los estudios de cine necesitan predecir qué tipo de películas es más probable que vayan a ver los aficionados. Mediante una regresión logística multinomial, el estudio puede determinar la influencia que la edad, el sexo y las relaciones de pareja de cada persona tienen sobre el tipo de película que prefieren. De esta manera, el estudio puede orientar la campaña publicitaria de una película concreta al grupo de la población que tenga más probabilidades de ir a verla.

Se obtienen estadísticos como el historial de iteraciones, coeficientes de los parámetros, covarianza asintótica y matrices de correlación, pruebas de la razón de verosimilitud para los efectos del modelo y los parciales, $-2 \log$ de la verosimilitud, chi-cuadrado de la bondad de ajuste de Pearson y de la desviación, R^{**2} de Cox y Snell, de Nagelkerke y de McFadden, frecuencias observadas respecto a las frecuencias pronosticadas por cada categoría de respuesta, tablas de contingencia para frecuencias observadas y pronosticadas (con los residuos) y proporciones por patrón en las covariables y por categoría de respuesta.

En cuanto a métodos, se ajusta un modelo logit multinomial para el modelo factorial completo o para un modelo especificado por el usuario. La estimación de los parámetros se realiza a través de un algoritmo iterativo de máxima verosimilitud.

Para realizar un análisis de regresión logística multinomial, elija en los menús *Analizar → Regresión → Logística binaria* (Figura 20-15) y seleccione las variables y las especificaciones para el análisis (Figura 20-16). Previamente es necesario cargar en memoria el fichero de nombre CRECIMIENTO mediante *Archivo → Abrir → Datos*. Este fichero contiene datos sobre características del organismo humano que son factores efectivamente interviniéntes en el crecimiento de los niños. La variable cualitativa dependiente va a ser el *índice* de crecimiento, el factor va a ser *género* (niño o niña) y las variables independientes son la distancia del centro de la pituitaria a la fisura ptérigo-maxilar (*distanci*) y la edad (*edad*). Se ajustará un modelo que prediga el crecimiento por sexo según las covariables.

En cuanto a los datos, la variable dependiente debe ser categórica. Las variables independientes pueden ser factores o covariables. En general, los factores deben ser variables categóricas y las covariables deben ser variables continuas.

En cuanto a los supuestos, se asume que la razón de ventajas de cualquier par de categorías es independiente de las demás categorías de respuesta. Según esta suposición, por ejemplo, si se introduce un nuevo producto en un mercado, la participación en el mercado de todos los demás productos quedará afectada de manera igualmente proporcional. De igual manera, dado un patrón en las covariables, se asume que las respuestas son variables multinomiales independientes.

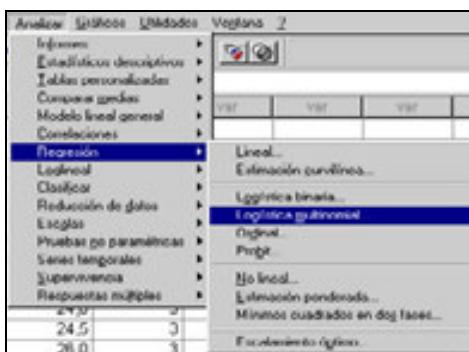


Figura 20-15



Figura 20-16

El botón *Modelos* (Figura 20-17) permite especificar diferentes modelos para una regresión logística multinomial. En *Especificar modelo* hay que tener en cuenta que un modelo de efectos principales contiene los efectos principales de las covariables y los factores, pero no contiene efectos de interacción. Un modelo factorial completo contiene todos los efectos principales y todas las interacciones factor por factor, pero no contiene interacciones de covariable. Puede crear un modelo personalizado para especificar subconjuntos de interacciones de los factores o bien interacciones de las covariables.

Si se elige *Personalizado*, el campo *Factores y Covariables* muestra una lista de los factores y las covariables, etiquetando con (F) los factores fijos y con (C) las covariables. El campo *Modelo* sirve para diseñar el modelo, que dependerá de los efectos principales y de interacción que seleccione. La opción *Incluir la intersección en el modelo* le permite incluir o excluir del modelo un término de intersección. El campo *Escala* permite especificar el valor de escalamiento de la dispersión que se va a utilizar para corregir la estimación de la matriz de covarianzas de los parámetros (*Desviación* estima el valor de escalamiento mediante el estadístico de la función de desviación chi-cuadrado de la razón de verosimilitud y *Pearson* estima el valor de escalamiento mediante el estadístico chi-cuadrado de Pearson). También puede especificar su propio valor de escalamiento especificando en el campo *Valor* un número positivo.

En el botón *Criterios* (Figura 20-18) puede especificar varios criterios para una regresión logística multinomial. *Iteraciones* le permite especificar el número máximo de veces que desea recorrer el algoritmo, el número máximo de pasos en la subdivisión por pasos, las tolerancias de convergencia para los cambios en el log de la verosimilitud y los parámetros y la frecuencia con que se imprime el progreso del algoritmo iterativo. *Delta* le permite especificar un valor no negativo inferior a 1. Este valor se añade a cada casilla vacía de la tabla de contingencia de las categorías de respuesta por patrones de covariables. Se ayuda así a estabilizar el algoritmo y evitar sesgos en las estimaciones. *Tolerancia para la singularidad* le permite especificar la tolerancia empleada en la comprobación de la singularidad. El botón *Estadísticos* lleva a las opciones de la Figura 20-19.

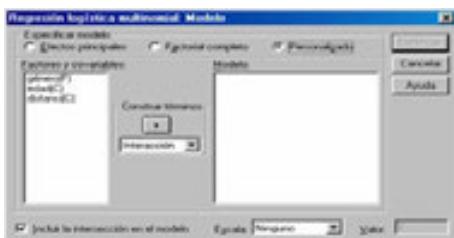


Figura 20-17

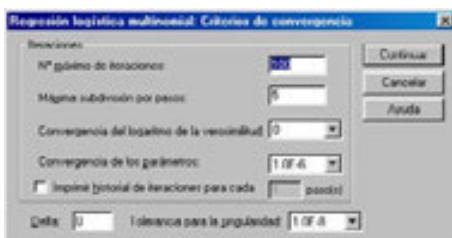


Figura 20-18

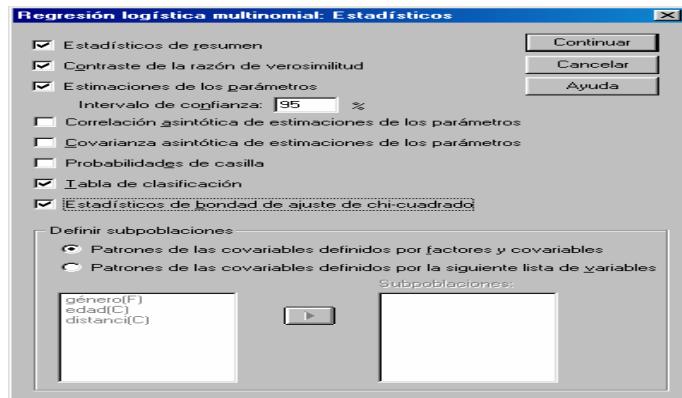


Figura 20-19

La opción *Guardar* permite exportar información del modelo al archivo especificado. SmartScore y las próximas versiones de What If? podrán utilizar este archivo. En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 20-16 para obtener los resultados del análisis según se muestra en la Figura 20-30. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 20-30 a 20-34 se presentan varias salidas de entre las múltiples que ofrece el procedimiento.

Resumen del procesamiento de los casos	
Índice	N
1	8
2	8
3	8
4	8
5	8
6	8
7	8
8	8
9	8
10	8
11	8
12	4
13	4
14	4
15	4
16	4
Género	
Niño	64
Niña	44
Válidos	108
Perdidos	0
Total	108

Figura 20-30

Información del ajuste del modelo				
Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	553,628			
Final	510,560	43,067	45	,554
Bondad de ajuste				
	Chi-cuadrado	gl	Sig.	
Pearson	986,424	1200	1,000	
Desviación	476,478	1200	1,000	
Pseudo R-cuadrado				
Cox y Snell	,329			
Nagelkerke	,330			
McFadden	,073			

Figura 20-31

Contrastes de la razón de verosimilitud					
Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.	
Intersección	510,560 ^a	,000	0	,	
EDAD	517,002 ^a	6,442	15	,971	
DISTANCI	529,627 ^a	19,066	15	,211	
GÉNERO	539,550	28,989	15	,016	

El estadístico de chi-cuadrado es la diferencia en las -2 log
verosimilitudes entre el modelo final y el modelo reducido. El modelo
reducido se forma omitiendo un efecto del modelo final. La hipótesis
nula es que todos los parámetros de ese efecto son 0.

^a Se han encontrado singularidades inesperadas en la matriz
Hessiana. Puede que los datos presenten una separación casi
completa. Algunas estimaciones de los parámetros tienden a
infinito.

Figura 20-32

Índice	B	Error Est.	VValid	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
							Límite inferior	Límite superior
1	Intersección	7,793	2654,167	,000	1	,999		
	EDAD	,693	,444	1,848	1	,174	,547	,229 1,305
	DISTANCI	,767	,374	4,210	1	,040	,154	1,035 4,482
	[GÉNERO=H]	-19,768	2654,161	,000	1	,994	2,549E-09	,000
	[GÉNERO=M]	0%			0			
2	Intersección	12,084	2654,166	,000	1	,999		
	EDAD	-,404	,441	,839	1	,360	,668	,281 1,585
	DISTANCI	-,446	,371	1,449	1	,230	,1,561	,754 3,232
	[GÉNERO=H]	-18,990	2654,161	,000	1	,994	5,650E-09	,000
	[GÉNERO=M]	0%			0			
3	Intersección	9,636	2654,167	,000	1	,997		
	EDAD	-,522	,442	1,398	1	,237	,593	,250 1,410
	DISTANCI	-,640	,373	2,947	1	,086	,1,896	,913 3,939
	[GÉNERO=H]	-19,465	2654,161	,000	1	,994	3,520E-09	,000
	[GÉNERO=M]	0%			0			
4	Intersección	3,453	2654,167	,000	1	,999		
	EDAD	-,782	,452	2,994	1	,084	,457	,189 1,109
	DISTANCI	1,033	,382	T,311	1	,007	,3,610	1,329 5,941
	[GÉNERO=H]	-20,537	2654,161	,000	1	,994	1,205E-09	,000
	[GÉNERO=M]	0%			0			

Figura 20-33

Observado	Pronosticado																Por con- taje cor- recto
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	0	0	1	3	3	0	0	0	0	0	0	0	0	1	0	,0%	
2	1	0	1	0	1	0	1	0	0	0	1	0	1	0	0	2	,0%
3	1	0	0	2	0	0	0	2	0	0	1	0	0	0	0	2	,0%
4	1	0	0	6	0	0	0	0	0	0	0	0	0	1	0	0	***
5	0	0	0	0	2	0	0	0	1	0	2	0	0	0	0	3	***
6	0	0	0	0	3	0	0	0	1	0	2	0	0	0	2	0	,0%
7	2	0	0	0	1	0	0	1	0	0	1	0	2	0	0	1	,0%
8	0	0	1	1	0	0	1	0	0	0	1	0	1	0	1	2	,0%
9	0	0	0	1	3	0	0	1	1	0	0	0	0	0	0	2	***
10	0	0	0	4	4	0	0	0	0	0	0	0	0	0	0	0	,0%
11	0	0	0	4	1	0	0	1	0	0	0	0	0	0	0	2	,0%
12	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	,0%
13	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	1	***
14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	***
15	0	0	2	1	0	0	0	0	1	0	0	0	0	0	0	0	,0%
16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	***
Porcentaje global	6%	,0%	6%	***	***	,0%	2%	6%	4%	,0%	7%	,0%	4%	2%	5%	***	

Figura 20-34

SPSS Y LOS MODELOS PROBIT Y LOGIT

SPSS incorpora un procedimiento que implementa el análisis de regresión probit. Este procedimiento mide la relación entre la intensidad de un estímulo y la proporción de casos que presentan una cierta respuesta a dicho estímulo. Es útil para las situaciones en las que se dispone de una respuesta dicotómica que se piensa puede estar influenciada o causada por los niveles de alguna o algunas variables independientes, y es particularmente adecuada para datos experimentales.

Este procedimiento le permitirá estimar la intensidad necesaria para que un estímulo llegue a inducir una determinada proporción de respuestas, como la dosis efectiva para la mediana. Como estadísticos se obtienen los coeficientes de regresión y errores típicos, intersección y su error típico, Chi-cuadrado de Pearson de la bondad de ajuste, frecuencias observadas y esperadas e intervalos de confianza para los niveles efectivos de la variable o variables independientes. Como diagramas se obtienen los gráficos de respuestas transformadas.

Como ejemplo podemos considerar la medición de la efectividad que tiene un nuevo pesticida para matar hormigas y cuál es la concentración adecuada que se debe utilizar. Para ello podría llevarse a cabo un experimento en el que se expongan muestras de hormigas a diferentes concentraciones del pesticida y después registrar el número de hormigas muertas y el número de hormigas expuestas. Aplicando el análisis probit a estos datos se puede determinar la fuerza de la relación entre concentración y mortalidad, así como determinar la concentración adecuada de pesticida si desea asegurar la extermación de, por ejemplo, el 95% de las hormigas expuestas.

Para realizar un *análisis probit*, elija en los menús *Analizar* → *Regresión* → *Probit* (Figura 20-35) y seleccione las variables y las especificaciones para el análisis (Figura 20-36). Previamente es necesario cargar en memoria el fichero de nombre EMPLEADOS mediante *Archivo* → *Abrir* → *Datos*. Este fichero contiene datos sobre el sexo de los trabajadores de una fábrica (*sexo*) clasificados según meses de experiencia previa (*expprev*) y sobre los que se ha medido su salario actual (*salario*) y su salario inicial (*salini*). Se ajustará un modelo probit que relacione la experiencia y el sexo de los trabajadores según la mejora de salario.

El botón *Frecuencia de respuesta* de la Figura 20-36 debe contener una variable de frecuencia de respuesta para la que cada caso debe contener el número de individuos que respondieron cuando fueron expuestos a un nivel del estímulo particular (o a una combinación de niveles estimulares, si dispone de varios estímulos). El botón *Total observado* debe contener una variable de frecuencia observada de modo que cada caso debe contener el número total de individuos expuestos a un nivel del estímulo o a una combinación de niveles estimulares.

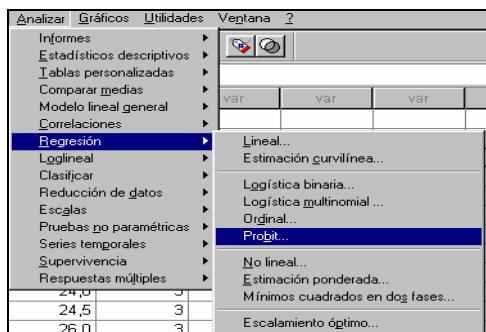


Figura 20-35

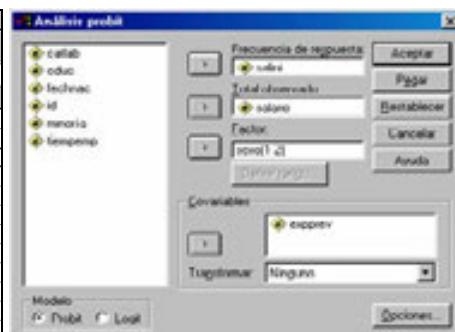


Figura 20-36

El botón *Factor* estima intersecciones diferentes para los subgrupos y debe indicar qué niveles del factor desea utilizar, mediante la definición del rango para el factor. En el campo *Covariables* introducimos otras variables para el modelo. Por defecto, las covariables no se transforman. De manera opcional, puede transformar las covariables eligiendo una transformación de la lista desplegable del botón *Transformar*. En el campo *Modelo* elegiremos *Probit* para aplicar la transformación Probit (la inversa de la función acumulada de la distribución normal típica) a las proporciones de respuesta, y elegiremos *Logit* para aplicar la transformación del *modelo Logit* (log de las ventajas) a las proporciones de respuesta.

El botón *Opciones* de la Figura 20-36 nos lleva a la Figura 20-37, en cuyos campos se pueden especificar opciones para el análisis probit. El campo *Estadísticos* permite solicitar los siguientes estadísticos opcionales: frecuencias, potencia relativa de la mediana, prueba de paralelismo e intervalos de confianza fiduciaria. Intervalos de confianza fiduciaria y potencia relativa de la mediana no están disponibles si se ha seleccionado más de una covariante. Potencia relativa de la mediana y prueba de paralelismo sólo están disponibles si se ha seleccionado una variable de factor. El campo *Tasa de respuesta natural* permite indicar una tasa de respuesta natural incluso en la ausencia del estímulo. Los posibles valores son Ninguna, Calcular a partir de los datos o Valor. El campo *Criterios* permite controlar los parámetros del algoritmo iterativo de estimación de los parámetros. Puede anular las opciones predeterminadas para N° máximo de iteraciones, Límite para los pasos y Tolerancia de la optimalidad.

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 20-36 para obtener los resultados del análisis según se muestra en la Figura 20-38. En la parte izquierda de la Figura podemos ir seleccionando los distintos tipos de resultados haciendo clic sobre ellos. También se ven los resultados desplazándose a lo largo de la pantalla. En las Figuras 20-38 a 20-42 se presentan varias salidas de entre las múltiples que ofrece el procedimiento.

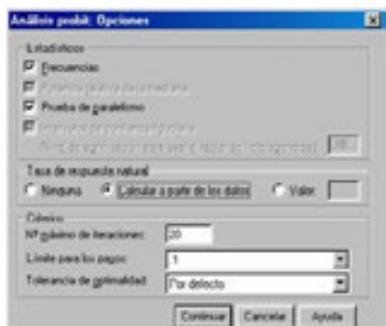


Figura 20-37

Figura 20-38

```
***** * PROBIT ANALYSIS *****

Parameter estimates converged after 6 iterations.
Optimal solution found.

Parameter Estimates (PROBIT model: (PROBIT(p)) = Intercept + BX):

Regression Coeff. Standard Error Coeff./S.E.

EXPPREV ,00100 ,000000 307,31272

Intercept Standard Error Intercept/S.E. SEXO
-,12625 ,000050 -250,05343 Hombre
-,06326 ,000058 -109,67552 Mujer

Pearson Goodness-of-Fit Chi Square = 512597,348 DF = 471 P = ,000
Parallelism Test Chi Square = 327,588 DF = 1 P = ,000

Since Goodness-of-Fit Chi square is significant, a heterogeneity
factor is used in the calculation of confidence limits.
```

Figura 20-39

Observed and Expected Frequencies						
SEXO	EXPPREV	Number of Subjects	Observed Responses	Expected Responses	Residual	Prob
1	144,00	57000,0	27000,0	28892,751	-1892,751	,50689
1	36,00	40200,0	18750,0	18652,685	97,315	,46400
1	138,00	45000,0	21000,0	22702,722	-1702,722	,50450
1	67,00	32100,0	13500,0	15288,852	-1788,852	,47629
1	114,00	36000,0	18750,0	17818,645	931,355	,49496
1	26,00	28350,0	12000,0	13042,107	-1042,107	,46004
1	34,00	27750,0	14250,0	12853,945	1396,055	,46321

Figura 20-40

Si en el campo *Modelo* de la Figura 20-36 elegimos la opción *Logit*, tenemos el ajuste de los datos a un modelo Logit (Figuras 20-41 y 20-42).

```
Parameter estimates converged after 6 iterations.
Optimal solution found.

Parameter Estimates (LOGIT model: (LOG(p/(1-p))) = Intercept + BX):

Regression Coeff. Standard Error Coeff./S.E.

EXPPREV ,00160 ,000001 306,02542

Intercept Standard Error Intercept/S.E. SEXO
-,20193 ,000081 -250,06536 Hombre
-,10131 ,000092 -109,92518 Mujer

Pearson Goodness-of-Fit Chi Square = 512466,180 DF = 471 P = ,000
Parallelism Test Chi Square = 324,368 DF = 1 P = ,000

Since Goodness-of-Fit Chi square is significant, a heterogeneity
factor is used in the calculation of confidence limits.
```

Figura 20-41

Observed and Expected Frequencies						
SEXO	EXPPREV	Number of Subjects	Observed Responses	Expected Responses	Residual	Prob
1	144,00	57000,0	27000,0	28900,634	-1900,634	,50703
1	36,00	40200,0	18750,0	18651,103	98,897	,46396
1	138,00	45000,0	21000,0	22708,470	-1708,470	,50463
1	67,00	32100,0	13500,0	15289,037	-1789,037	,47629
1	114,00	36000,0	18750,0	17821,715	928,285	,49505
1	26,00	28350,0	12000,0	13040,632	-1040,632	,45999
1	34,00	27750,0	14250,0	12852,780	1397,220	,46316

Figura 20-42

PROCEDIMIENTO TABLAS DE MORTALIDAD

Para crear una tabla de mortalidad en SPSS, elija en los menús *Analizar* → *Supervivencia* → *Tablas de mortalidad* (Figura 20-43) y seleccione una variable numérica de supervivencia en el campo *tiempo* de la Figura 20-44. Consideraremos el fichero LEUCEMIA que contiene datos sobre el número de semanas de tratamiento (*tiempo*) de pacientes a los que se somete o no a quimioterapia (*quimio*) reflejando el resultado de la misma (*estado*) y siendo el evento terminal el valor 1 de la variable *estado*, es decir, la recaída. Se generarán tablas de mortalidad para pacientes que han recibido quimioterapia y para los que no la han recibido.

En el campo *Mostrar intervalos de tiempo* especifique los intervalos de tiempo que se van a examinar (extremo superior de todos los intervalos en la casilla *0 hasta* y anchura de los intervalos en la casilla *por*). En el campo *Estado* seleccione una variable de estado para definir casos para los que tuvo lugar el evento terminal. Pulse en *Definir evento* para especificar el valor de la variable de estado, el cual indica que el evento ha tenido lugar (Figura 20-45). Si lo desea, en el campo *Factor* puede seleccionar una variable de factor de primer orden e introducir su rango con el botón *Definir rango* (Figura 20-46). Se generan tablas actuariales de la variable de supervivencia para cada categoría de la variable de factor. Además es posible seleccionar una variable por factor de segundo orden en el campo *Por factor*. Las tablas actuariales de la variable de supervivencia se generan para cada combinación de las variables de factor de primer y segundo orden. Pulse en el botón *Opciones* (Figura 20-47) para generar tablas, gráficos y comparar niveles de factor.

En cuanto a los datos la variable de tiempo deberá ser cuantitativa. La variable de estado deberá ser dicotómica o categórica, codificada en forma de números enteros, con los eventos codificados en forma de un valor único o un rango de valores consecutivos. Las variables de factor deberán ser categóricas, codificadas como valores enteros. Se supone que las probabilidades para el evento de interés deben depender solamente del tiempo transcurrido desde el evento inicial (se asume que son estables con respecto al tiempo absoluto). Es decir, los casos que se introducen en el estudio en horas diferentes (por ejemplo, pacientes que inician el tratamiento en horas diferentes) se deberían comportar de manera similar. Tampoco deben existir diferencias sistemáticas entre los casos censurados y los no censurados. Si, por ejemplo, muchos de los casos censurados son pacientes en condiciones más graves, los resultados pueden resultar sesgados.

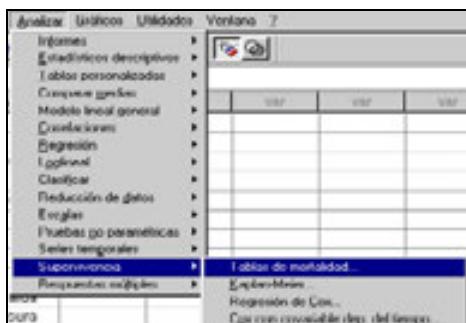


Figura 20-43

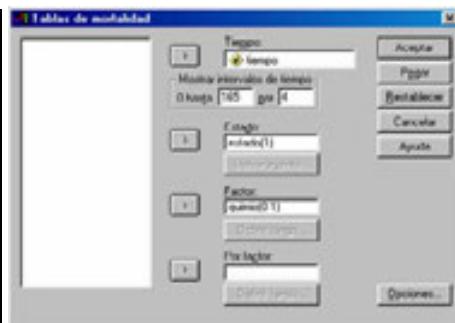


Figura 20-44



Figura 20-45



Figura 20-46

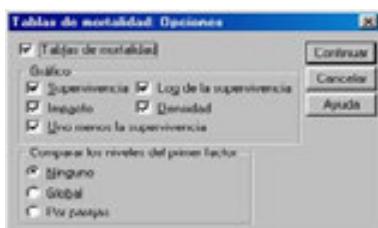


Figura 20-47

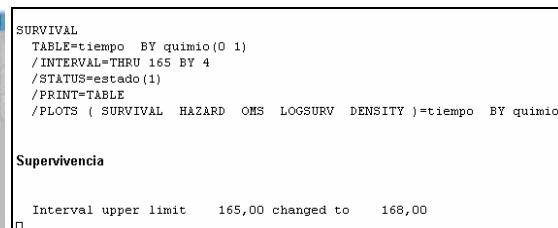


Figura 20-48

This subfile contains: 23 observations

Life Table

Survival Variable	TIEMPO	Tiempo (semanas)
	for QUIMIO	Quimioterapia de mantenimiento
		= 0 No

Intrvl Start	Number Entrng	Number Wdrawn	Number Exposd	Number of	Propn Termi-	Propn Sur-	Cumul	Probabi-	Hazard
Time Intrvl	Intrvl	Intrvl	Risk	Events	nating	viving	Propn at End	Density	Rate
,0	12,0	,0	12,0	,0	,0000	1,0000	1,0000	,0000	,0000
4,0	12,0	,0	12,0	2,0	,1667	,8333	,8333	,0417	,0455
8,0	10,0	,0	10,0	2,0	,2000	,8000	,6667	,0417	,0556
12,0	8,0	,0	8,0	1,0	,1250	,8750	,5833	,0208	,0333
16,0	7,0	1,0	6,5	,0	,0000	1,0000	,5833	,0000	,0000
20,0	6,0	,0	6,0	1,0	,1667	,8333	,4861	,0243	,0455
24,0	5,0	,0	5,0	1,0	,2000	,8000	,3889	,0243	,0556
28,0	4,0	,0	4,0	1,0	,2500	,7500	,2917	,0243	,0714
32,0	3,0	,0	3,0	1,0	,3333	,6667	,1944	,0243	,1000
36,0	2,0	,0	2,0	,0	,0000	1,0000	,1944	,0000	,0000
40,0	2,0	,0	2,0	1,0	,5000	,5000	,0972	,0243	,1667
44,0	1,0	,0	1,0	1,0	,0000	,0000	,0243	,0000	

The median survival time for these data is 23,43

Figura 20-49

Life Table Survival Variable TIEMPO Tiempo (semanas) for QUIMIO Quimioterapia de mantenimiento = 1 Sí										
Start Time	Intrvl	Number Entrng	Number Udrawn	Number Exposd	Number of Events	Propn Termnl	Propn Termini-	Cumul Propn	Proba-	Hazard Rate
	Intrvl	Risk					Surviving	Surv	Density	
,0	,0	11,0	,0	11,0	,0	,0000	1,0000	1,0000	,0000	,0000
4,0	,0	11,0	,0	11,0	,0	,0000	1,0000	1,0000	,0000	,0000
8,0	,0	11,0	,0	11,0	1,0	,0909	,9091	,9091	,0227	,0238
12,0	,0	10,0	,0	9,5	1,0	,1053	,8947	,8134	,0239	,0278
16,0	,0	8,0	,0	8,0	1,0	,1250	,8750	,7117	,0254	,0333
20,0	,0	7,0	,0	7,0	1,0	,1429	,8571	,6100	,0254	,0385
24,0	,0	6,0	,0	6,0	,0	,0000	1,0000	,6100	,0000	,0000
28,0	,0	6,0	1,0	5,5	1,0	,1818	,8182	,4991	,0277	,0500
32,0	,0	4,0	,0	4,0	1,0	,2500	,7500	,3743	,0312	,0714
36,0	,0	3,0	,0	3,0	,0	,0000	1,0000	,3743	,0000	,0000
40,0	,0	3,0	,0	3,0	,0	,0000	1,0000	,3743	,0000	,0000
44,0	,0	3,0	1,0	2,5	,0	,0000	1,0000	,3743	,0000	,0000
48,0	,0	2,0	,0	2,0	1,0	,5000	,5000	,1872	,0468	,1667
52,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000
56,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000
60,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000
64,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000
68,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000
72,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000
76,0	,0	1,0	,0	1,0	,0	,0000	1,0000	,1872	,0000	,0000

Figura 20-50

Intrvl Start Time	SE of Cumul Surviving	SE of Proba- bility	SE of Hazard Rate
,0	,0000	,0000	,0000
4,0	,1076	,0269	,0320
8,0	,1361	,0269	,0390
12,0	,1423	,0199	,0333
16,0	,1423	,0000	,0000
20,0	,1481	,0230	,0453
24,0	,1470	,0230	,0552
28,0	,1387	,0230	,0707
32,0	,1219	,0230	,0980
36,0	,1219	,0000	,0000
40,0	,0919	,0230	,1571
44,0	,0000	,0230	,0000

Figura 20-51

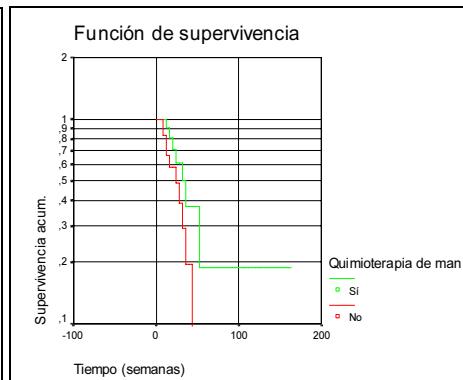


Figura 20-52

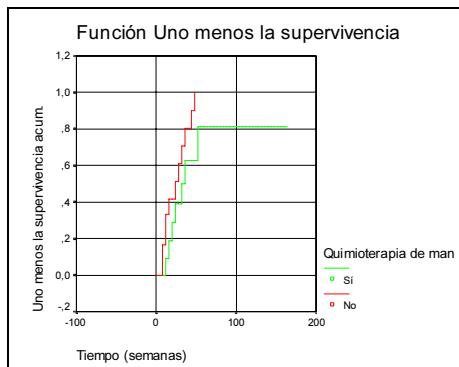


Figura 20-53

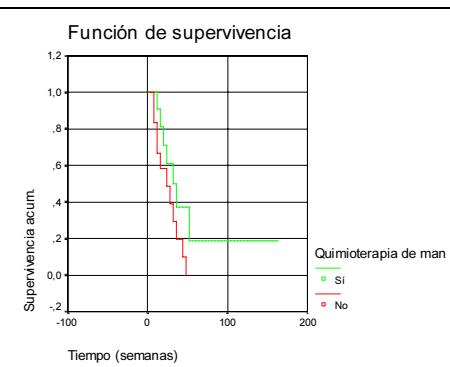


Figura 20-54

Al pulsar *Aceptar* en la Figura 20-44 se obtiene la salida del procedimiento con su sintaxis (Figura 20-48), las tablas de mortalidad para cada nivel de factor (Figuras 20-49 y 20-50), los valores de las funciones de densidad, supervivencia y azar (Figura 20-51) y sus gráficos (Figuras 20-52 a 20-54).

PROCEDIMIENTO KAPLAN-MEIER

Para obtener un análisis de supervivencia de Kaplan-Meier con SPSS, elija en los menús *Analizar* → *Supervivencia* → *Kaplan-Meier* (Figura 20-55) y seleccione una variable de tiempo en el campo *Tiempo* de la Figura 20-56. Seleccione una variable de estado que identifique los casos para los que ha tenido lugar el evento terminal en el campo *Estado*. Esta variable puede ser numérica o de cadena corta. Pulse en *Definir evento* para especificar el valor de la variable de estado que indica que el evento ha tenido lugar (Figura 20-57). Si lo desea, puede seleccionar una variable de factor para examinar las diferencias entre grupos en el campo *Factor*. Además es posible seleccionar una variable de estrato, que generará análisis diferentes para cada nivel (cada estrato) de la variable. Al igual que en el procedimiento anterior usaremos el fichero LEUCEMIA y sus variables. En el botón *Comparar niveles de los factores* (Figura 20-58) se pueden solicitar estadísticos para contrastar la igualdad de las distribuciones de supervivencia para los diferentes niveles del factor. En el botón *Opciones* (Figura 20-59) se eligen estadísticos y gráficos a mostrar y en el botón *Guardar* (Figura 20-60) se eligen resultados a guardar. Al pulsar *Aceptar* se obtiene la salida (Figuras 20-61 a 20-67).

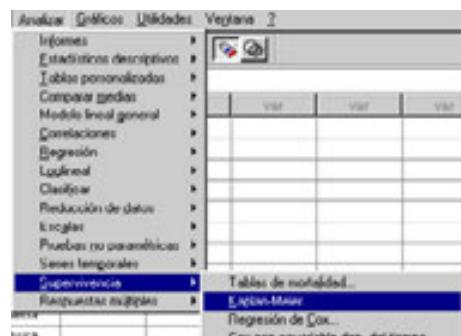


Figura 20-55

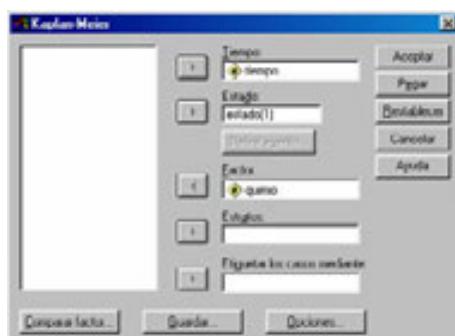


Figura 20-56



Figura 20-57

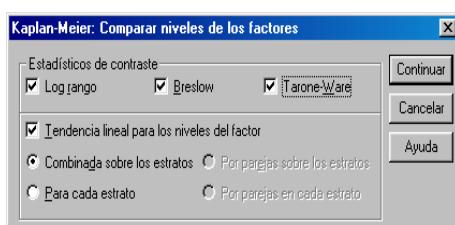


Figura 20-58

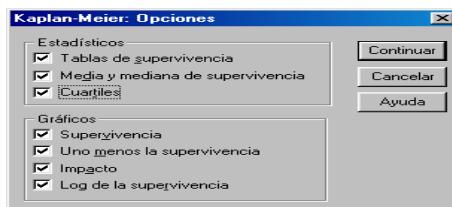


Figura 20-59

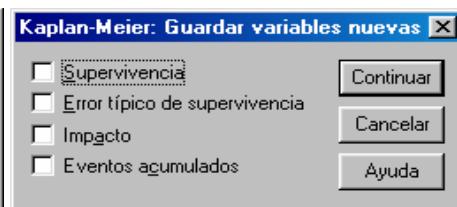


Figura 20-60

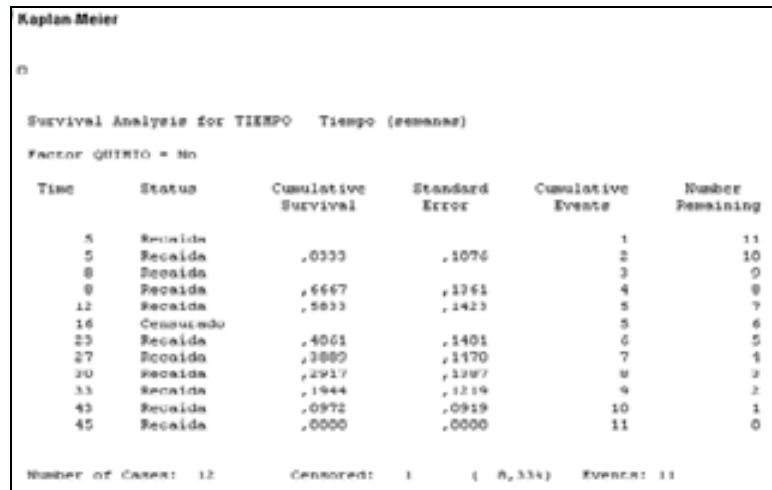


Figura 20-61

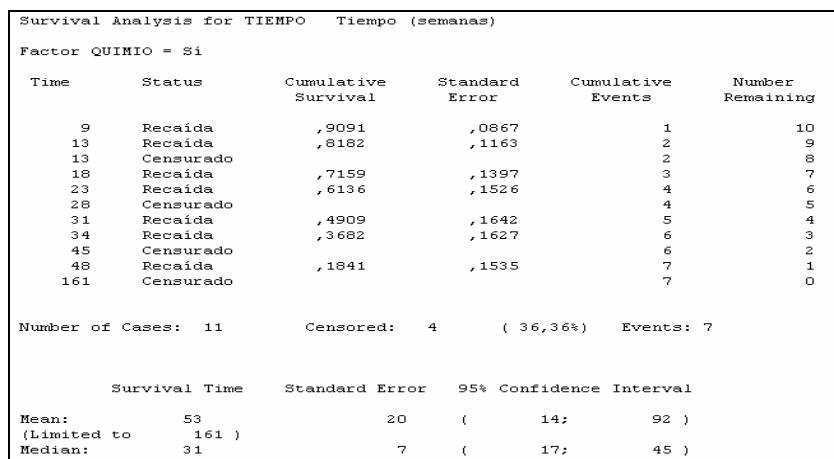


Figura 20-62

Survival Analysis for TIEMPO Tiempo (semanas)					
		Total	Number Events	Number Censored	Percent Censored
QUIMIO	No	12	11	1	8,33
QUIMIO	Si	11	7	4	36,36
Overall		23	18	5	21,74

>Note # 20103. Command name: KM
>Since no metric was specified on the TREND subcommand the default is used.

Test Statistics for Equality of Survival Distributions for QUIMIO with Trend, metric = (-1, 1)

Statistic	df	Significance
Log Rank	3,40	,0653
Breslow	2,72	,0989
Tarone-Ware	2,98	,0842

Figura 20-63

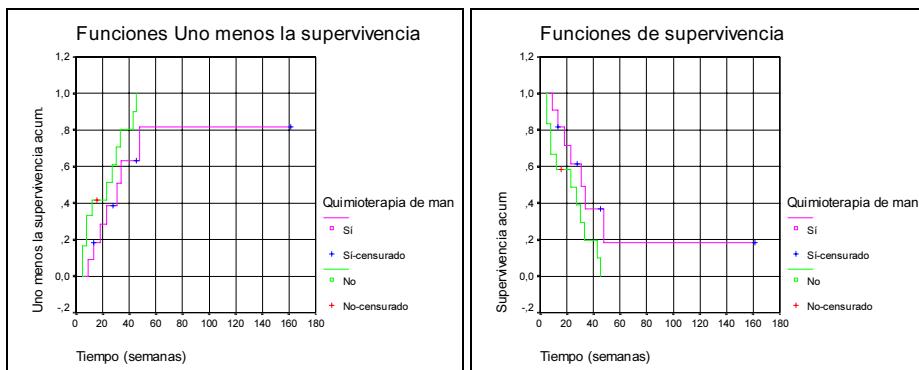


Figura 20-64

Figura 20-65

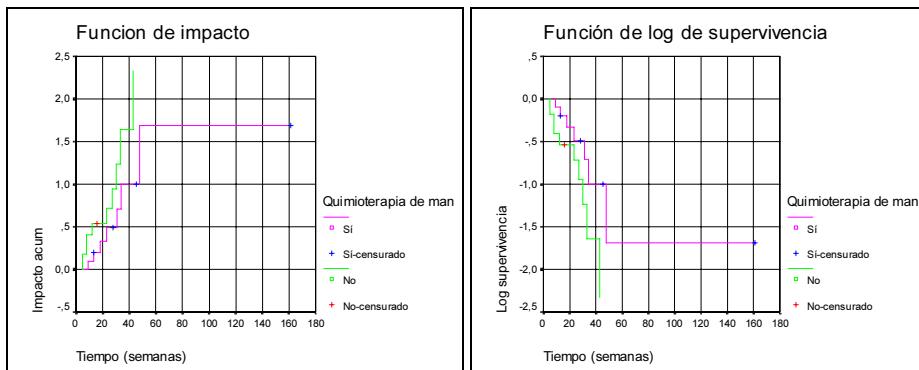


Figura 20-66

Figura 20-67

Se han representado las tablas de mortalidad para cada nivel de factor (Figuras 20-61 y 20-62), los estadísticos de igualdad de distribuciones de supervivencia (Figura 20-63) para cada nivel de factor (con p-valores no muy lejanos a 0,05 que indican proximidad o lejanía débil entre las distribuciones) y las gráficas de las funciones de riesgo, supervivencia, azar e impacto (Figuras 20-64 a 20-67).

PROCEDIMIENTO REGRESIÓN DE COX

Del mismo modo que las tablas de mortalidad y el análisis de supervivencia de Kaplan-Meier, la regresión de Cox es un método para crear modelos para datos de tiempos de espera hasta un evento con casos censurados presentes. Sin embargo, la regresión de Cox permite incluir en los modelos variables predictoras (covariables). Por ejemplo, a partir de los datos del fichero EMPLEADOS podrá construir un modelo de la duración en el empleo (*tiempemp*) como función del nivel educativo (*educ*) y de la categoría laboral (*catlab*). La regresión de Cox gestionará los casos censurados correctamente y proporcionará las estimaciones de los coeficientes para cada una de las covariables, permitiendo evaluar el impacto de múltiples covariables en el mismo modelo. Además, es posible utilizar la regresión de Cox para examinar el efecto de covariables continuas como por ejemplo el *salario*.

Como estadísticos se obtienen para cada modelo: -2LL, el estadístico de la razón de verosimilitud y el Chi-cuadrado global. Para las variables dentro del modelo: Estimaciones de los parámetros, errores típicos y estadísticos de Wald. Para variables que no estén en el modelo: estadísticos de puntuación y Chi-cuadrado residual. En cuanto a los datos, la variable de tiempo debería ser cuantitativa y la variable de estado puede ser categórica o continua. Las variables independientes (las covariables) pueden ser continuas o categóricas; si son categóricas, deberán ser auxiliares (*dummy*) o estar codificadas como indicadores (existe una opción dentro del procedimiento para recodificar las variables categóricas automáticamente). Las variables de estratos deberían ser categóricas, codificadas como valores enteros o cadenas cortas.

Para obtener un análisis de regresión de Cox, elija en los menús *Analizar* → *Supervivencia* → *Regresión de Cox* (Figura 20-68). En la Figura 20-69 seleccione una variable para el campo *Tiempo*, seleccione una variable de *Estado* y pulse en *Definir evento* (Figura 20-70) para caracterizar el valor o rango de valores que determinan el evento. Seleccione variables para utilizarlas como *Covariables*. Si lo desea, es posible calcular modelos diferentes para diferentes grupos definiendo una variable para los *Estratos*. En el botón *Categórica* (Figura 20-71) se pueden introducir variables categóricas como covariables. En el botón *Guardar* (Figura 20-72) se definen los resultados a guardar en el botón *Opciones* (Figura 20-73), se definen distintas características a obtener en la regresión y en el botón *Gráficos* se definen los distintos gráficos a obtener (Figura 20-74). Al pulsar *Aceptar* se obtiene la salida con el ajuste del modelo y los gráficos (Figuras 20-75 a 20-80).

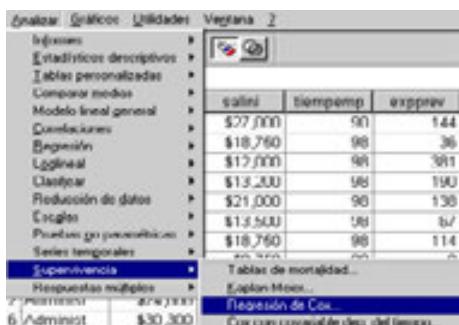


Figura 20-68



Figura 20-69



Figura 20-70

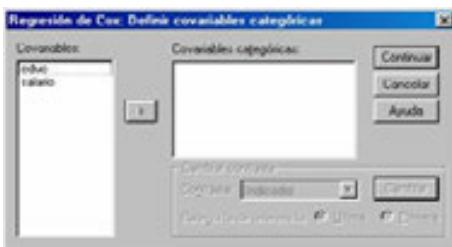


Figura 20-71

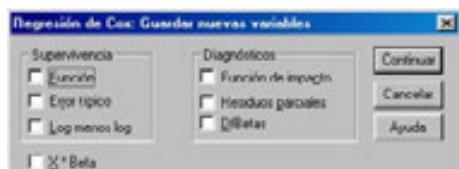


Figura 20-72

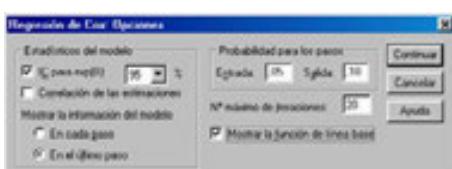


Figura 20-73

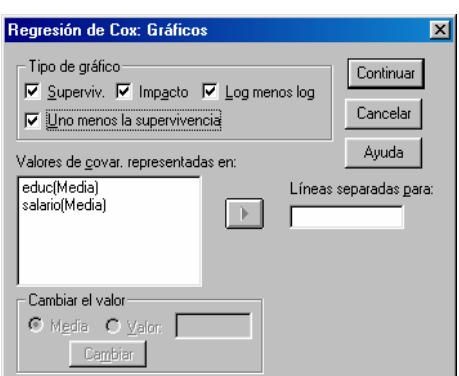


Figura 20-74

Resumen del proceso de casos		
	N	Porcentaje
Casos disponibles en el análisis	474	100,0%
Evento*	0	,0%
Censurado	0	,0%
Total	474	100,0%
Casos excluidos		
Casos con valores perdidos	0	,0%
Casos con tiempo no positivo	0	,0%
Casos censurados antes del evento más temprano en un estrato	0	,0%
Total	0	,0%
	474	100,0%

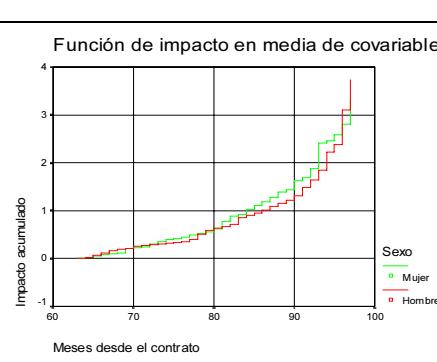
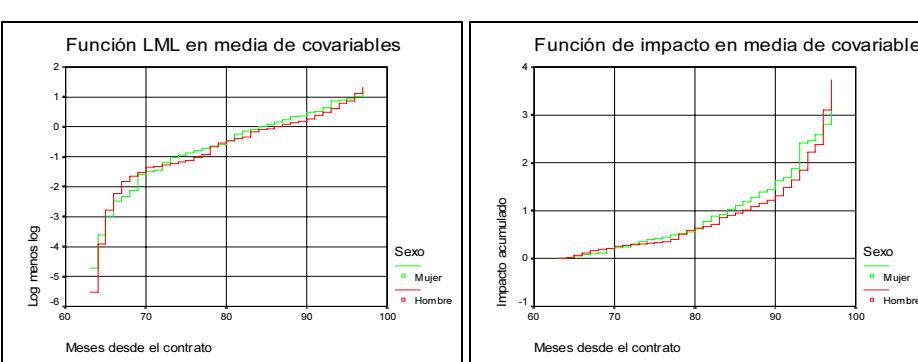
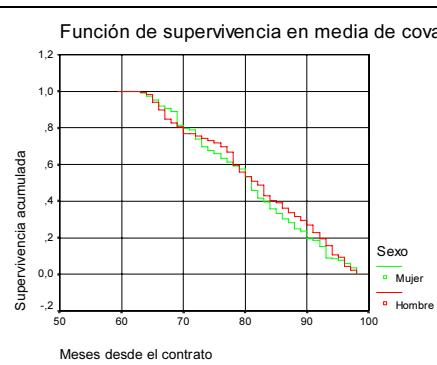
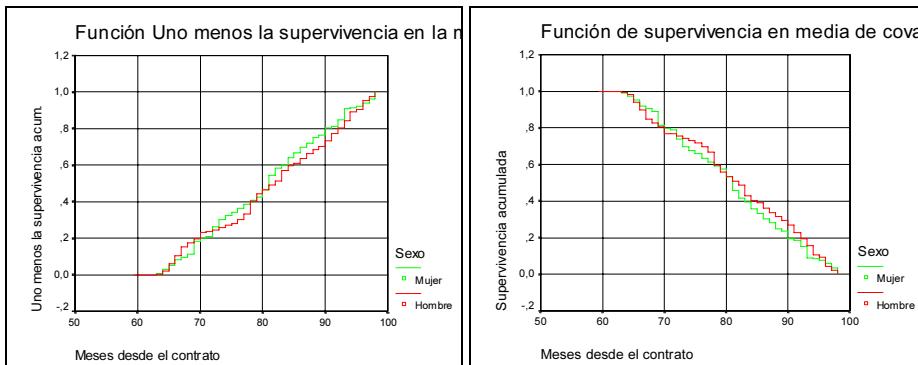
Figura 20-75

Pruebas omnibus sobre los coeficientes del modelo									
-2 log de la verosimilitud									
4324,410									
Bloque 1: Método = Introducir									
Pruebas omnibus sobre los coeficientes del modelo ^{a,b}									
-2 log de la verosimilitud	Global (puntuación)			Cambio desde el paso anterior			Cambio desde el bloque anterior		
	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
4322,383	1,938	2	,379	2,028	2	,363	2,028	2	,363

a. Bloque inicial número 0, función log de la verosimilitud inicial: -2 log de la verosimilitud: 4324,410
b. Bloque inicial número 1. Método = Introducir

Variables en la ecuación									
	B	ET	Wald	gl	Sig.	Exp(B)	95,0% IC para Exp(B)		
							Inferior	Superior	
EDUC	-,002	,022	,008	1	,928	,998	,957	1,041	
SALARIO	,000	,000	1,144	1	,285	1,000	1,000	1,000	

Figura 20-76



En SPSS existe una variante de la regresión de Cox, que consiste en utilizar como covariable una variable dependiente del tiempo definida a medida por el usuario. El procedimiento de SPSS se denomina **Regresión de Cox con covariable dependiente del tiempo** y su pantalla de entrada (Figura 20-81) se utiliza para definir la citada covariable en función de la variable T_ que aporta SPSS de forma automática al elegir el procedimiento. El botón *Modelo* nos lleva a la Figura 20-82 donde se definen las variables del modelo como en el procedimiento anterior. Al pulsar *Aceptar* se obtiene la salida (Figuras 20-83 a 20-85).



Figura 20-81

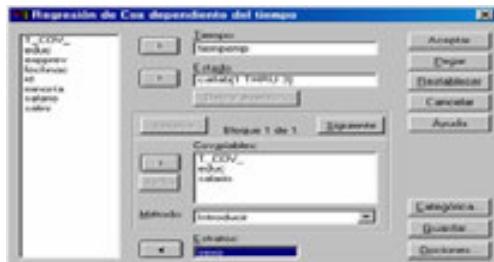


Figura 20-82

Resumen del proceso de casos		
	N	Porcentaje
Casos disponibles	474	100,0%
en el análisis		
Censurado	0	,0%
Total	474	100,0%
Casos incluidos		
Casos con valores	0	,0%
perdidos		
Casos con tiempo no	0	,0%
nádiles		
Casos censurados	0	,0%
antes del evento más		
temprano en un estrato		
Total	0	,0%
Total	474	100,0%

B. Variable dependiente: Meses desde el contrato.

Estado del estrato ^a			
Estrato	Número de estrato	Evento	Porcentaje (estrato)
1	Mujeres	258	0
2	Hombres	216	0
Total		474	0

^a La variable de estrato es: Sexo

Figura 20-83

Bloque 0: Bloque inicial			
Pruebas omnibus sobre los coeficientes del modelo			
$-2 \log de la verosimilitud^a$			
4324,410			
Bloque 1: Método = Introducir			
Historial de iteraciones ^b			
	$-2 \log de la verosimilitud^a$	Coefficiente	
	T_COV_ EDUC SALARIO	T_COV_ EDUC SALARIO	
1	3178,431	-,242 ,012 ,000	
2	2715,885	-,504 ,004 ,000	
3	2375,089	-,926 ,002 ,000	

^a Bloque inicial número 0, función log de la verosimilitud inicial: -2 log de la verosimilitud: 4324,410

^b Al menos un coeficiente tiende al infinito después de 3 iteraciones

Figura 20-84

Pruebas omnibus sobre los coeficientes del modelo ^a									
-2 log de la verosimilitud	Global (puntuación)			Cambio desde el paso anterior			Cambio desde el bloque anterior		
	Chi-cua drado	gl	Sig.	Chi-cua drado	gl	Sig.	Chi-cua drado	gl	Sig.
2375,089	953,700	3	,000	1949,321	3	,000	1949,321	3	,000

^a Bloque inicial número 1. Método = Introducir

Variables en la ecuación									
B	ET	VValid	gl	Sig.	Exp(B)	95,0% IC para Exp(B)	Inferior	Superior	
T_COV_	-,926	,048	366,108	1	,000	,396	,360	,436	
EDUC	,002	,022	,010	1	,921	1,002	,960	1,046	
SALARIO	,000	,000	,001	1	,981	1,000	1,000	1,000	

Medias de las covariables	
	Media
T_COV_	11,698
EDUC	13,563
SALARIO	35173,967

Figura 20-85

Ejercicio 20-1. Se considera una muestra de 53 pacientes con cáncer de próstata en los que se mide la edad, el nivel de ácido que mide la extensión del tumor, el grado de agresividad del tumor, la etapa en la que se encuentra, los resultados de una radiografía y cuándo se ha detectado al intervenir quirúrgicamente que el cáncer se ha extendido a los nodos linfáticos. A partir de estos datos se trata de ajustar un modelo que permita predecir cuándo el cáncer se extiende a los nodos linfáticos sin necesidad de intervención quirúrgica.

Rellenamos la pantalla de entrada del procedimiento *Regresión Logística Binaria* tal y como se indica en la Figura 20-86 y se rellena la pantalla del botón *Opciones* como se indica en la Figura 20-87. Al pulsar *Continuar* y *Aceptar* se obtiene la salida que incluye las Figuras 20-88 y 20-89.



Figura 20-86

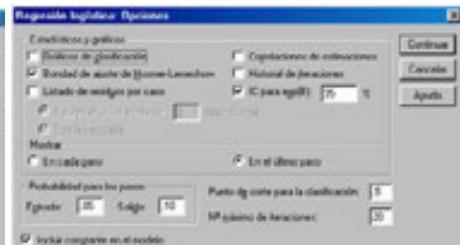


Figura 20-87

Pruebas omnibus sobre los coeficientes del modelo				
		Chi-cuadrado	gl	Sig.
Paso 1	Paso	22,126	5	,000
	Bloque	22,126	5	,000
	Modelo	22,126	5	,000

Resumen de los modelos				
	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke	
1	48,126	,341	,465	

Prueba de Hosmer y Lemeshow				
	Chi-cuadrado	gl	Sig.	
1	5,954	8	,652	

Figura 20-88

Variables en la ecuación								
	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
1 ^a	ACIDO	,024	,013	3,423	1	,064	1,025	,999 1,051
	EDAD	-,069	,058	1,432	1	,231	,933	,833 1,045
	ETAPA	1,564	,774	4,083	1	,043	4,778	1,048 21,783
	GRADO	,761	,771	,976	1	,323	2,141	,473 9,700
	RADIOGRA	2,045	,807	6,421	1	,011	7,732	1,589 37,614
	Constante	,062	3,460	,000	1	,986	1,064	

a. Variable(s) introducida(s) en el paso 1: ACIDO, EDAD, ETAPA, GRADO, RADIOGRA.

Figura 20-89

Los estadísticos y p-valores de las Figuras 20-88 y 20-89 revelan buen ajuste y significatividad para el modelo, así como intervalos de confianza para los parámetros. La ecuación de ajuste ha resultado ser la siguiente:

$$p = \frac{1}{1 + e^{-(0,062 + 0,024 \text{ÁCIDO} - 0,069 \text{EDAD} + 1,564 \text{ETAPA} + 0,761 \text{GRADO} + 2,045 \text{RADIOGRAFÍA})}}$$

Para un hombre de 66 años con un nivel de ácido de 48 y con valor cero para el resto de las variables, tenemos que la probabilidad de que el cáncer se extienda a los nodos linfáticos es:

$$p = \frac{1}{1 + e^{-(3,346)}} = 0,0340$$

Ejercicio 20-2. Se dispone de tres tipos de pesticida (rotenone, deguelin y mixture) que al ser aplicados en diferentes dosis sobre un número total de insectos provocan la muerte de una cierta cantidad de ellos. Mediante un modelo Probit se trata de hallar la relación entre tipo de pesticida y su efecto por muerte en los insectos.

Rellenamos la pantalla de entrada del procedimiento *Regresión Probit* tal y como se indica en la Figura 20-90 y se rellena la pantalla del botón *Opciones* como se indica en la Figura 20-91. Al pulsar *Continuar* y *Aceptar* se obtiene la salida que incluye las Figuras 20-92 y 20-93.

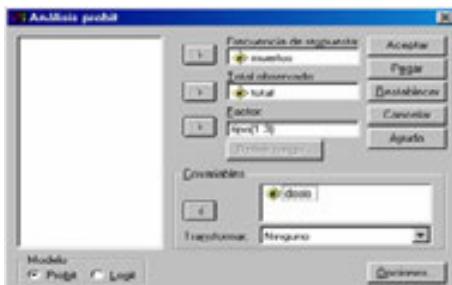


Figura 20-90

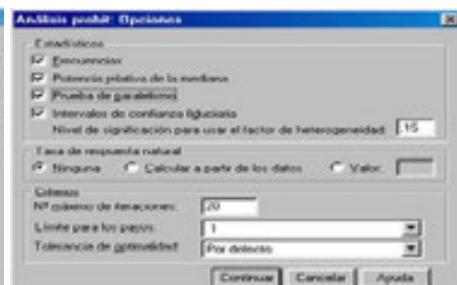


Figura 20-91

Figura 20-92

```
***** PROBIT ANALYSIS *****

Parameter estimates converged after 13 iterations.
Optimal solution found.

Parameter Estimates (PROBIT model: (PROBIT(p)) = Intercept + BX):

Regression Coeff. Standard Error Coeff./S.E.

DOSIS ,10235 ,01048 9,76567

Intercept Standard Error Intercept/S.E. TIPO
-,48731 ,10295 -4,73341 rotenone
-1,54637 ,23297 -6,63769 deguelin
-,74473 ,19776 -3,76585 mixture

Pearson Goodness-of-Fit Chi Square = 76,971 DF = 10 P = ,000
Parallelism Test Chi Square = 63,441 DF = 2 P = ,000

Since Goodness-of-Fit Chi square is significant, a heterogeneity
factor is used in the calculation of confidence limits.
```

Figura 20-93

Según estos resultados, los modelos para cada tipo de pesticida son:

$$\text{Probit}(P_i) = -0,487 + 0,102(\text{dosis}_i)$$

$$\text{Probit}(P_i) = -1,546 + 0,102(\text{dosis}_i)$$

$$\text{Probit}(P_i) = -0,744 + 0,102(\text{dosis}_i)$$

Ejercicio 20-3. Consideramos los tiempos de remisión de una enfermedad (en semanas) en 42 pacientes, 12 de ellos censurados (señalados con un signo +), que son los siguientes:

1, 10, 22, 7, 3, 32+, 12, 23, 8, 22, 17, 6, 2, 16, 11, 34+, 8, 32+, 12, 25+, 2, 11+, 5, 20+, 4, 19+, 15, 6, 8, 17+, 23, 35+, 5, 6, 11, 13, 4, 9+, 1, 6+, 8, 10+

Construir la tabla de mortalidad y las funciones de azar, densidad y supervivencia para este modelo. Representar la función de supervivencia.

Comenzamos introduciendo los datos de los tiempos de remisión en una variable en el editor de SPSS con nombre REM. Adicionalmente introducimos otra variable de nombre C1 cuyos valores son 1 o 0 según que los datos de REM sean censurados o no. Rellenamos la pantalla de entrada del procedimiento *Tablas de mortalidad* tal y como se indica en la Figura 20-94, y se rellena la pantalla del botón *Opciones* como se indica en la Figura 20-95. Al pulsar *Continuar* y *Aceptar* se obtiene la salida que incluye las Figuras 20-96 a 20-98.

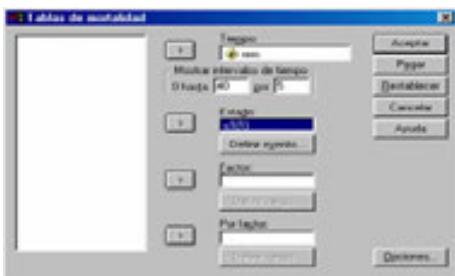


Figura 20-94

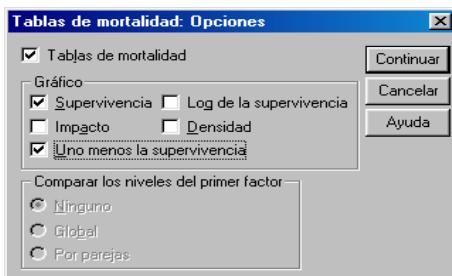


Figura 20-95

```
SURVIVAL TABLE=rem / INTERVAL=THRU 40 BY 5 / STATUS=c1(1) / PRINT=TABLE
/PLOTS ( SURVIVAL OMS )=rem .
```

Supervivencia

□

This subfile contains: 62 observations

Life Table
Survival Variable REM

Intrvl Start	Number Entred this	Number Withdrawn	Number Exposed	Number of Events	Propn Termnl	Propn Estimating	Cumul Propn	Probabi-ty	Hazard Rate
Time	Intrvl	Intrvl	Risk	Events	Termnl	Estimating	Surv at End	Density	
5,0	42,0	7,0	38,5	,0	,0000	,0000	,0000	,0000	,0000
5,0	35,0	10,0	30,0	,0	,067	,033	,033	,013	,013
10,0	23,0	6,0	20,0	,0	,1000	,0000	,0400	,0187	,0211
15,0	15,0	3,0	13,5	,0	,1481	,0519	,7156	,0249	,0320
20,0	10,0	4,0	8,0	,0	,1250	,0750	,6261	,0179	,0267
25,0	5,0	,0	5,0	,0	,2000	,0000	,5009	,0250	,0444
30,0	4,0	,0	4,0	,0	,7500	,2500	,1252	,0751	,2400
35,0	1,0	,0	1,0	,0	1,0000	,0000	,0000	,0250	,4000

The median survival time for these data is 30,01

Figura 20-96

Intrvl Start	SE of Cumul	SE of Probabi-	SE of Hazard
Time	Sur-viving	lity	Rate
,0	,0000	,0000	,0000
5,0	,0455	,0014	,087
10,0	,0748	,0126	,0149
15,0	,1032	,0164	,0226
20,0	,1231	,0169	,0266
25,0	,1492	,0229	,0442
30,0	,1147	,0312	,1109
35,0	,0000	,0229	,0000

Figura 20-97

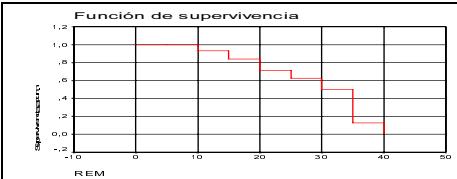


Figura 20-98

Ejercicio 20-4. En la tabla siguiente se presenta un estudio sobre la evolución de 62 pacientes que presentan cáncer de pulmón. Se conoce el tiempo de observación en días, el estado de fallecimiento (1) o vida (0) al final del período de observación y el tipo de tratamiento administrado (1 = estándar y 0 = experimental).

Tiempo	Estado	Tipo	Tiempo	Estado	Tipo	Tiempo	Estado	Tipo
72	0	1	112	0	0	3	0	1
411	0	1	999	0	0	95	0	1
228	0	1	11	0	1	24	0	0
231	1	0	25	1	1	18	0	0
242	0	0	144	0	1	83	1	0
991	0	0	8	0	1	31	0	0
111	0	0	42	0	1	51	0	0
1	0	0	100	1	1	90	0	0
587	0	0	314	0	1	52	0	0
389	0	0	110	0	1	73	0	0
33	0	0	82	0	1	8	0	0
25	0	0	10	0	1	36	0	0
357	0	0	118	0	1	48	0	0
467	0	0	126	0	1	7	0	0
201	0	0	8	0	1	140	0	0
1	0	0	92	0	1	186	0	0
30	0	0	35	0	1	84	0	0
44	0	0	117	0	1	19	0	0
283	0	0	132	0	1	45	0	0
15	0	0	12	0	1	80	0	0
87	1	0	162	0	1			

Realizar la estimación no paramétrica de las funciones de supervivencia para los dos tratamientos simultáneamente y graficarlas en los mismos ejes. Comparar las tablas de vida y ajustar una regresión de Cox con la covariante tipo como categórica.

Comenzamos introduciendo los datos de los tiempos de observación, estado y tipo en tres variables de nombres TOBS, EST y TIP respectivamente, como tres columnas del editor de datos de SPSS. Rellenamos la pantalla de entrada del procedimiento *Kaplan-Meier* tal y como se indica en la Figura 20-99. y se rellena la pantalla del botón *Opciones* como se indica en la Figura 20-100 y el botón *Comparar factor* como se indica en la Figura 20-101. Al pulsar *Continuar* y *Aceptar* se obtiene la salida que incluye las tablas de supervivencia para cada tipo (Figuras 20-102 y 20-103), los estadísticos de comparación de tablas (Figura 20-104) y las funciones de supervivencia e impacto (Figuras 20-105 y 20-106).

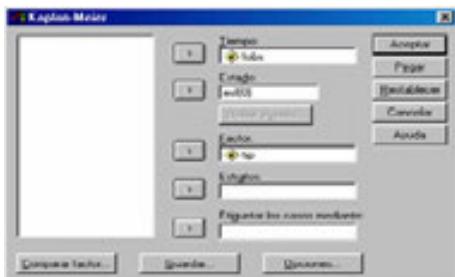


Figura 20-99

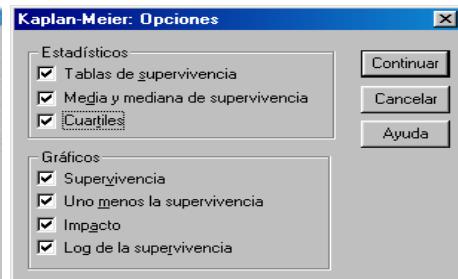


Figura 20-100



Figura 20-101

Survival Analysis for TOBS						
Factor TIP = 0						
Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining	
1	0	,9474	,0362	1	37	
1	0	,9211	,0437	2	36	
7	0	,8947	,0498	3	35	
8	0	,8684	,0548	4	34	
15	0	,8421	,0592	5	33	
18	0	,8158	,0629	6	32	
19	0	,7895	,0661	7	31	
24	0	,7632	,0690	8	30	
25	0	,7368	,0714	9	29	
30	0	,7105	,0736	10	28	
31	0	,6842	,0754	11	27	
33	0	,6579	,0770	12	26	
36	0	,6316	,0783	13	25	
44	0	,6053	,0793	14	24	
45	0			15	23	
Survival Time Standard Error 95% Confidence Interval						
Mean:	188	45	(99;	278)	
Median:	73	25	(24;	122)	
Percentiles						
	25,00	50,00	75,00			
Value	242,00	73,00	30,00			
Standard Error	98,10	25,06	7,60			

Figura 20-102

Survival Analysis for TOBS					
Factor TIP = 1					
Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
3	0	,9583	,0408	1	23
8	0	,8750	,0675	2	22
8	0	,8750	,0675	3	21
10	0	,8333	,0761	4	20
11	0	,7917	,0829	5	19
12	0	,7500	,0884	6	18
25	1			6	17
35	0	,7059	,0936	7	16
42	0	,6618	,0976	8	15
72	0	,6176	,1005	9	14
82	0	,5735	,1026	10	13
92	0	,5294	,1037	11	12
95	0	,4853	,1041	12	11
100	1			12	10
110	0	,4368	,1044	13	9
117	0	,3882	,1034	14	8
118	0	,3397	,1012	15	7
126	0	,2912	,0977	16	6
132	0	,2426	,0927	17	5
144	0	,1941	,0859	18	4
162	0	,1456	,0769	19	3
228	0	,0971	,0648	20	2
314	0	,0485	,0472	21	1
411	0	,0000	,0000	22	0

Number of Cases:	24	Censored:	2	(8,33%)	Events:	22
 Survival Time Standard Error 95% Confidence Interval						
Mean: 111 22 (67; 155)						
Median: 95 21 (53; 137)						
 Percentiles						
25,00 50,00 75,00						
Value 132,00 95,00 12,00						
Standard Error 16,56 21,30 16,49						
 Survival Analysis for TOBS						
Total Number Events Number Censored Percent Censored						
TIP 0	38	35	3	7,89		
TIP 1	24	22	2	8,33		
Overall	62	57	5	8,06		

Figura 20-103

Test Statistics for Equality of Survival Distributions for TIP			
	Statistic	df	Significance
Log Rank	,77	1	,3803
Breslow	,01	1	,9229
Tarone-Ware	,15	1	,6978

Figura 20-104

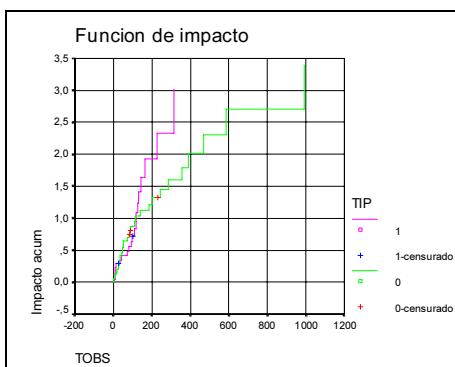


Figura 20-105

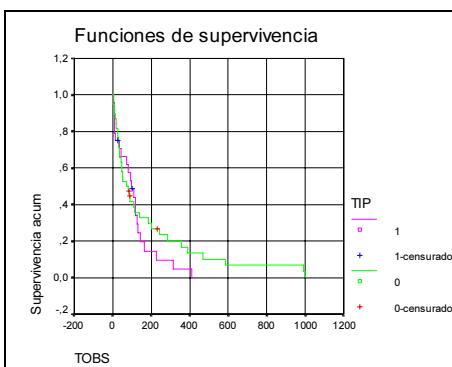


Figura 20-106

Para ajustar los datos a una regresión de Cox con tipo como covariante categórica, rellenamos la pantalla de entrada del procedimiento *Regresión de Cox* como se indica en la Figura 20-107. Las pantallas de los botones *Categoría* y *Opciones* se llenan como se indica en las Figuras 20-108 y 20-109. Al pulsar *Continuar* y *Aceptar* se obtiene el ajuste de las Figuras 20-110 y 20-111.



Figura 20-107

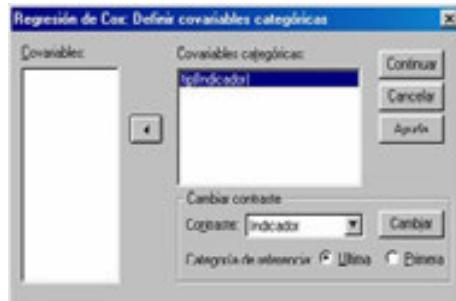


Figura 20-108

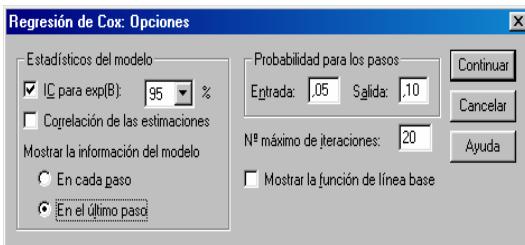


Figura 20-109

Codificación de variables categóricas ^{a,b}		
	Frecuencia	(1)
TIP	0	,38
	1	,000 .000

a. Codificación de parámetros de indicador
b. Variable de categorías: TIP

Bloque 0: Bloque inicial

Pruebas omnibus sobre los coeficientes del modelo

-2 log de la verosimilitud	361,333
----------------------------	---------

Figura 20-110

Bloque 1: Método = Introducir									
Pruebas omnibus sobre los coeficientes del modelo ^{a,b}									
-2 log de la verosimilitud	Global (puntuación)			Cambio desde el paso anterior			Cambio desde el bloque anterior		
	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
360,581	,768	1	,381	,753	1	,386	,753	1	,386

a. Bloque inicial número 0, función log de la verosimilitud inicial: -2 log de la verosimilitud: 361,333
b. Bloque inicial número 1. Método = Introducir

Variables en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)	95,0% IC para Exp(B)
TIP	-,247	,283	,764	1	,382	,781	Inferior ,448 Superior 1,360

Medias de las covariantes

	Media
TIP	,613

Figura 20-111

ANÁLISIS CONJUNTO

CONCEPTO DE ANÁLISIS CONJUNTO

Uno de los problemas básicos de la investigación de mercados es la necesidad de descubrir qué características de un producto o servicio son más importantes para los consumidores. Esto es importante sobre todo para el diseño de nuevos productos, la reposición de productos existentes, la evaluación de los efectos de precios en la decisión de hacer una compra, la simulación de cuotas de mercado, etc.

El Análisis de Conjunto es, precisamente, una técnica estadística que determina qué características de un producto (o servicio) son las preferidas por los consumidores y cuantifica estas preferencias. Las características de un producto incluyen atributos como la marca, el color, formas, precio y garantía y el análisis de conjunto mide las preferencias del consumidor por las características particulares de un producto.

El análisis de conjunto se basa en la suposición de que los consumidores toman la decisión de compra considerando simultáneamente todas las características del producto. Para hacer esto, los consumidores deben de buscar un equilibrio en términos de la relación calidad-precio, porque normalmente un producto no tiene todas las mejores características. Por ejemplo, el típico coche grande de lujo proporciona un mayor status, seguridad y tamaño, pero también cuesta más y tiene mayor consumo por kilómetro. El análisis de conjunto se utiliza para estudiar estos equilibrios.

El Análisis de Conjunto compite con los tradicionales métodos de investigación que a menudo preguntan a los consumidores sobre la importancia de cada característica de un producto por separado.

El análisis de conjunto es superior a estos otros métodos ya que se basa en modelos más precisos de la forma en que los consumidores toman decisiones analizando los descartes que hay entre las características, proporciona un mejor indicador de la importancia relativa de una característica en contraste con los métodos tradicionales en los que los consumidores tendían a evaluar todas las características de un producto como importantes para la decisión de la compra, proporciona investigaciones con un producto hipotético, con una particular combinación de características y niveles, que son los preferidos por los consumidores. Esto es muy útil en el desarrollo de productos, pues un nuevo producto puede ser desarrollado en base a este producto hipotético.

Desde mediados de los años 70 el análisis conjunto ha atraído una atención considerable tanto como método que representa de modo realista las decisiones de los consumidores como por los descartes entre multiatributos de productos o servicios. Ha ganado una amplia aceptación y uso en muchas industrias con tasas de utilización que aumentaron hasta diez veces más en los años ochenta. Esta aceleración ha coincidido con la amplia introducción de programas de ordenador que integran el proceso entero, generando las combinaciones de los valores de la variable predictora para ser evaluadas con la creación de simuladores de opciones que predicen elecciones del consumidor a través de un gran número de formulaciones de productos o servicios alternativos. El acceso a estos programas ha aumentado aún más con la introducción de programas basados en ordenadores personales. Hoy, cualquier investigador con un ordenador personal puede acceder a varios paquetes ampliamente utilizados.

El análisis conjunto está altamente relacionado con la experimentación tradicional. Por ejemplo, un químico en una planta de fabricación de jabón puede querer saber el efecto que en las tinajas de fabricación del jabón tienen la temperatura y la presión sobre la densidad de la pastilla de jabón resultante. El químico podría conducir un experimento de laboratorio para medir esas relaciones. Una vez dirigidos los experimentos, podrían analizarse con procedimientos ANOVA (Análisis de la Varianza).

El análisis conjunto es una técnica muy importante para comprender las reacciones de los consumidores y las evaluaciones de las combinaciones de atributos predeterminados que representan potenciales productos o servicios. Mantiene un alto grado de realismo y proporciona al analista una mejor comprensión de la composición de las preferencias del cliente. La flexibilidad del análisis conjunto proviene de su capacidad para acomodarse tanto a una variable dependiente métrica como a una no métrica, del uso de variables predictoras categóricas y de las muchas asunciones generales acerca de las relaciones de las variables independientes con la variable dependiente.

El análisis conjunto es una técnica multivariante utilizada específicamente para entender cómo los encuestados desarrollan preferencias para productos o servicios. Se basa en la sencilla premisa de que los encuestados evalúan el valor o utilidad de un producto/servicio/idea (real o hipotética) procedente de la combinación de las cantidades separadas de utilidad suministradas por cada atributo.

El análisis conjunto es el único de entre todos los métodos multivariantes en el que el investigador construye primero un conjunto de productos o servicios reales o hipotéticos por combinación de los niveles seleccionados de cada atributo. Estos productos hipotéticos se presentan más tarde a los encuestados que suministran únicamente sus evaluaciones globales. Así, el investigador está preguntando al encuestado para desempeñar una tarea tan realista como es la opción entre un conjunto de productos. Los encuestados no necesitan decir nada más al investigador que lo importante que es un producto para ellos o lo bien que el producto representa un número de atributos. Dado que el investigador construyó los hipotéticos productos/servicios de forma específica, la importancia de cada atributo y de cada valor de cada atributo pueden determinarse por los prorratores globales de los encuestados.

Para tener éxito, el analista debe ser capaz de describir el producto o servicio tanto en términos de sus atributos como en el de todos los valores importantes para cada atributo. Utilizaremos el término *factor* cuando estemos describiendo un atributo específico u otra característica del producto o servicio. Los valores posibles para cada factor se llaman *niveles*. En términos conjuntos, describimos un producto o servicio en base a *su nivel sobre el conjunto de factores que lo caracterizan*. Cuando el analista selecciona los factores y los niveles para describir un producto/servicio conforme a un plan específico, la combinación se llama *tratamiento o estímulo*.

EL ANÁLISIS CONJUNTO COMO UNA TÉCNICA MULTIVARIANTE DE LA DEPENDENCIA

El análisis conjunto puede expresarse en términos de modelo de dependencia como sigue:

$$\underbrace{Y_1}_{\text{No métrica o métrica}} = \underbrace{X_1 + X_2 + \cdots + X_N}_{\text{No métricas}}$$

El análisis conjunto es una técnica estadística utilizada para analizar la relación lineal o no lineal entre una variable dependiente (o endógena) generalmente ordinal (aunque también puede ser métrica) y varias variables independientes (o exógenas) no métricas.

La expresión funcional del análisis conjunto puede escribirse también como sigue:

$$y = F(x_1, x_2, \dots, x_n)$$

La variable dependiente recoge la preferencia (intención de compra, etc.) que el individuo exhibe hacia el producto (es decir, la utilidad global que el producto le aporta) y las variables dependientes son los atributos distintivos del producto.

Es importante tener presente que sólo la variable dependiente recogerá información aportada por los individuos encuestados, ya que la información contenida en las variables independientes será especificada por el investigador en virtud de los productos que desee someter a evaluación por los encuestados.

El análisis conjunto permite generar un modelo individualizado por encuestado, de modo que el modelo general para toda la muestra resulte de la agregación de los modelos de todos los individuos que la componen. El análisis conjunto descompone las preferencias que el individuo manifiesta hacia el producto a fin de conocer qué valor le asigna a cada atributo (*técnica descomposicional*), mientras que en el análisis discriminante y en el análisis de la regresión las valoraciones de cada atributo que hace el sujeto se utilizan para componer su preferencia sobre el producto (*técnicas composicionales*).

TÉCNICAS COMPOSICIONALES Y DESCOMPOSICIONALES

Cualquier estímulo (producto, marca o servicio) es percibido por múltiples atributos que, además, son evaluados de manera compensatoria. Más concretamente, un sujeto puede preferir un estímulo con déficit en un atributo porque este déficit se puede compensar con el resto de atributos. Así pues, los estímulos son multiatributos y, por tanto, las preferencias o el juicio asociado a un estímulo será el resultado del efecto conjunto de las características o atributos que definen el estímulo.

No obstante, partiendo del supuesto de compensación entre los atributos, las respuestas de preferencia de los sujetos pueden analizarse siguiendo un enfoque composicional o descomposicional.

La metodología composicional está más relacionada con los modelos sintéticos en los que los sujetos evalúan individualmente las diferentes características de los estímulos.

En la metodología descomposicional el sujeto analiza perfiles de atributos o estímulos globales.

Cuando en investigación deseamos modelizar y explicar las preferencias de los sujetos (consumidores), podemos recurrir a cualquiera de las dos metodologías que, por otra parte, deben ser consideradas complementarias e independientes. Sin embargo, la estrategia descomposicional es la que representa una mayor validez en cuanto que los sujetos tienen que evaluar los estímulos tal y como se los encuentra en la vida real. Por otra parte, el modelo descomposicional permite, durante la recogida de datos, que los sujetos puedan llevar a cabo tareas más complejas e informativas a través de las cuales pueden poner en marcha las reglas compensatorias anteriormente asumidas durante los procesos del conocimiento y el comportamiento.

Una característica importante de la metodología descomposicional es que se basa en supuestos más realistas y, además recurre a procesos indirectos a partir de los juicios de proximidad o preferencia que dan los sujetos, tanto para la identificación de los atributos como para conocer su peso.

Una característica importante de la metodología composicional o autoexplicativa, es que en ella se obtienen juicios directos sobre la importancia relativa de los atributos. Lo ideal sería combinar ambos tipos de datos ya que así se sumaría la riqueza y naturalidad de los juicios conjuntos con los bajos costes temporales y de dificultad para emitir juicios autoexplicativos.

Para la metodología composicional existe el programa ACA (*Adaptative Conjoint Análisis*) que permite un método de recogida de datos interactivo para la obtención de los datos autoexplicativos y conjuntos. Sin embargo se ha demostrado que los modelos tradicionales o puramente descomposicionales y los modelos híbridos presentan la misma capacidad predictiva que los modelos composicionales.

Para la metodología descomposicional se dispone del algoritmo CONJOINT, que será el utilizado en este libro.

El algoritmo CONJOINT refleja el proceso general del método de Análisis Conjunto, comenzando por la identificación y selección de las variables más relevantes en el estudio concreto y continuando por una serie de fases características entre las que se incluyen cómo formular el plan experimental y cómo obtener diseños factoriales fraccionados ortogonales, la elección de la estrategia para obtener la información deseada por parte del consumidor, la elección del tipo de modelos a suponer entre las respuestas de los sujetos y los valores de utilidad, el decidir entre métodos métricos o no métricos a la hora de seleccionar un algoritmo de estimación de utilidades, la decisión de cómo evaluar los resultados, la valoración de la importancia relativa de las variables predictoras y de cada uno de sus niveles en la influencia de los juicios del consumidor, la aplicación de un simulador de opciones a los resultados conjuntos para la predicción de los juicios del consumidor sobre nuevas combinaciones de atributos, y otras muchas tareas.

APLICACIONES DEL ANÁLISIS CONJUNTO

Una aplicación general de la utilización del análisis conjunto podría ser el conocimiento de la importancia que tiene una determinada característica o atributo en la decisión global de preferencia del sujeto hacia ese producto o servicio.

En las aplicaciones del análisis conjunto se emplea la lógica del diseño de experimentos y se utiliza el ajuste de modelos lineales a variables ordinales. Se asume un modelo del comportamiento multiatributo según el cual los sujetos tienen la capacidad de percibir cada uno de los atributos que configuran un estímulo.

Hay muchas disciplinas en la que se aplica el análisis conjunto. Por ejemplo, en Marketing se utiliza el análisis conjunto en análisis de la demanda o conducta de compra a partir de los modelos multiatributo. Autores como Carroll, Fishbein, Kruskal y Rosenberg., Varela, J., y Braña, T. han trabajado en esta vertiente del análisis conjunto.

Otra materia con grandes aplicaciones del análisis conjunto es la Investigación Comercial. El análisis conjunto es útil siempre que se deseen identificar las actitudes de los consumidores en la decisión de compra. Por lo tanto, será una herramienta útil en el análisis dinámico de la demanda de productos o servicios.

Ya sabemos que con la utilización del análisis conjunto se asume un modelo del comportamiento multiatributivo según el cual los sujetos tienen la capacidad de percibir cada uno de los atributos que configuran un estímulo. Por lo tanto, el análisis conjunto será útil en Psicología Comercial, siempre que se desee explorar y cuantificar el sistema de valores de los sujetos en el momento de elegir una alternativa entre varias posibles. En esta disciplina hay evidencia empírica de que las preferencias del consumo de bienes y servicios responden directamente a una percepción evaluativa de sus atributos y no a una percepción global.

Ya hemos dicho que con la aplicación del análisis conjunto podremos conocer qué importancia tiene una determinada característica o atributo en la decisión global de preferencia del sujeto hacia ese producto. Sin embargo, será necesario clarificar el concepto de atributo. Desde el punto de vista psicológico, por atributo se entiende una propiedad extraída de la experiencia humana. Es, pues, una propiedad que atribuimos a alguna cosa y no la cosa misma. Los atributos se relacionan con la experiencia personal o percepción que tenemos de una característica o propiedad del producto. En la metodología conjunta se distingue entre la dimensión del objeto físico (o característica) y la percepción de dicha característica (o atributo). Como ejemplo, el sonido tiene una intensidad física, pero también podemos hablar de la sonoridad que experimentamos. En cuanto al sabor, éste se puede medir mediante la cantidad de cloruro sódico o glucosa (característica), pero también podemos medir la sensación de sabor dulce o salado que experimenta el sujeto (atributo).

Tanto en Marketing como en Psicología, lo que suele interesar no es tanto la medición del estímulo externo o característica, sino el atributo subyacente que se experimenta evaluando la sensación o significado que tiene para el consumidor. Lo esencial en la práctica es que manipulando convenientemente los atributos se incide en las actitudes hacia ese producto o servicio. Los atributos se utilizan entonces para configurar actitudes o reacciones del consumidor hacia los productos o estímulos.

Ya hemos citado anteriormente que, desde el punto de vista de recogida de datos, los modelos multiatributo se agrupan en composicionales (o de síntesis) y descomposicionales. En los primeros se pregunta a los consumidores por la utilidad de los diferentes niveles de atributos considerados individualmente o por pares y, a continuación, se estima la utilidad global del producto. El algoritmo más conocido para este caso es el TRADE-OFF, implementado en el paquete estadístico PCMDS. Por otra parte, los modelos descompositivos o analíticos, lo que pretenden es estimar la aportación de los diferentes niveles de atributos en la construcción de las preferencias globales a partir de preguntas reflejadas al perfil completo del producto real.

Para aplicaciones en Marketing, los modelos descompositivos parecen más apropiados y tienen mayor apoyo empírico por presentar mayor capacidad predictiva de la intención o elección de compra. Por esta razón, muchos autores definen al análisis conjunto como un método descomposicional.

En cuanto a las aplicaciones del análisis conjunto en la Investigación Comercial, los pioneros fueron Green y Rao (1971), que definieron el análisis conjunto como «un conjunto de técnicas y modelos que buscan transformar las respuestas subjetivas de los consumidores en parámetros que sirvan para estimular la utilidad de cada nivel de atributo en la respuesta de preferencia manifestada por los consumidores.

El análisis conjunto es un instrumento esencial para modelizar las preferencias, hecho que está ampliamente justificado por las teorías psicológicas basadas en el procesamiento de la información, las cuales mantienen que las personas realizamos una evaluación individual de los productos o marcas en todas sus dimensiones o atributos y la utilidad total de un producto vendrá dada por la suma de las utilidades de cada atributo. Pero además, estos modelos de decisión multiatributo son compensatorios, dado que las utilidades de los atributos son aditivas. En este sentido, el comportamiento de compra se explicará a partir del producto que obtenga la puntuación más alta.

Una manera muy sencilla de presentar formalmente el análisis conjunto podría ser a través de la formalización que Anderson (1974) presenta formalmente el análisis conjunto como la *teoría de la integración de la información*.

Según Anderson, los juicios de preferencia (Y) se pueden expresar como una función entre las características de los estímulos (X) y un conjunto de coeficientes (C) que ponderan la aportación de cada información parcial o atributo a la información total [Y = f(C, X)]. El análisis conjunto nos va a permitir estimar y conocer esos coeficientes que modelan las propiedades de los estímulos.

Tanbién el análisis conjunto puede utilizarse en la Teoría de la Decisión, materia en la que los trabajos siguen un modelo de elección en el que los sujetos tienden a elegir una opción entre varias y, generalmente, se olvidan que los objetos y estímulos sobre los hay que tomar decisiones son complejos ya que pueden descomponerse en diferentes atributos susceptibles de ser evaluados de manera diferente. El enfoque del análisis conjunto se centra en el estudio de las propiedades del estímulo y, de esta manera, se considera que puede ser un perfecto complemento al estudio de las variables propias del sujeto o grupo, ampliamente tratadas en la teoría de la decisión.

El uso del análisis conjunto está recomendado siempre que busquemos el desarrollo de un diseño eficaz del producto con éxito. Sabemos que en cualquier situación de decisión el consumidor se enfrenta a una elección entre diferentes productos conformados por una serie de características que pueden tomar distintos valores. Por ejemplo, ante la elección de una carrera universitaria se pueden considerar como atributos la ciudad donde se imparte, sus diferentes salidas profesionales, dificultad o fracaso escolar, conocimientos previos exigidos, etc. Así mismo, dentro de las ciudades donde se imparte pueden existir diferentes lugares o niveles: Santiago, León, Sevilla y Madrid, por ejemplo. El análisis conjunto nos va a permitir modelizar las decisiones de los estudiantes y así poder diseñar la oferta ideal desde el punto de vista de éstos, sin embargo, antes hemos de responder a preguntas como: ¿qué atributos del producto o servicio son importantes y cuáles no para el consumidor o usuario?, ¿qué niveles concretos de atributos son los preferidos en la mente del consumidor?, ¿cuál es la cuota de mercado prevista de nuestra oferta real o simulada?, etc. Además, si disponemos de información adicional de los sujetos (respuestas a otras preguntas de la encuesta), entonces podremos identificar los segmentos del mercado a quienes hacer llegar los distintos servicios o productos. Por ejemplo, ¿el estudiante con perfil de «letras» y el de «ciencias» podrían tener distintas preferencias que podrían suscitar distintas ofertas por nuestra parte?

Ya sabemos que para medir el valor o utilidad que le da el consumidor o usuario a cada uno de los niveles de los atributos de un producto o servicio, existen los métodos de balance, composicionales o autoexplicados (que tratan de calcular esa utilidad preguntando directamente al sujeto por cada uno de los niveles), y los métodos descompositivos o métodos de balance de múltiples factores (las utilidades se estiman a partir de la opinión del individuo acerca de una serie de perfiles globales combinación de niveles de atributos). El análisis conjunto (CONJOINT) es un método típicamente descomposicional.

Desde que en 1978, Green y Srinivasan utilizan por primera vez el análisis conjunto para describir las preferencias del consumidor en el ámbito de la economía y el marketing, se ha ido concretando el tipo de preguntas a las que da respuesta la utilización de la medición conjunta. La mayoría de estas preguntas están orientadas a saber qué características del producto son importantes para los clientes, cuáles son los niveles de características preferidos y cómo realizar con eficacia estudios de precios en equilibrios con la marca.

ANÁLISIS CONJUNTO A TRAVÉS DEL PERFIL COMPLETO

No olvidemos que el análisis conjunto es un método que nos permite determinar qué características de un producto o servicio son las preferidas por los consumidores y cuantificar esas preferencias mediante utilidades. Posteriormente, esas utilidades pueden ser utilizadas para la toma de decisiones en el campo del diseño de nuevos productos o servicios, rediseño de productos ya existentes, evaluación de los efectos del precio en la decisión de hacer una compra, simulación de cuotas de mercado, etc. En definitiva, podemos decir que la utilización del análisis conjunto contribuye al estudio de los procesos de decisión a dos niveles: por un lado, como herramienta para modelizar las preferencias y, por otro, como recordatorio a la teoría de la decisión de la necesidad de incorporar los estímulos y no centrarse únicamente en los sujetos.

La mayoría de software estadístico, y en concreto SPSS (software que utilizaremos en el próximo capítulo) utiliza la aproximación de perfil completo (*full profile*) para aplicar el análisis conjunto. En este caso los sujetos que responden a la encuesta o manipulan directamente los estímulos, elaboran un rango de los perfiles o estímulos alternativos definidos por los niveles particulares de todos los atributos estudiados.

Esta aproximación de perfil completo requiere echar mano de los *diseños factoriales fraccionados*, ya que nos proporcionan una fracción adecuada en todas las alternativas posibles, lo que nos permite obtener datos fiables al tiempo que reducir la dificultad de la tarea. Para generar un diseño ortogonal de efectos principales de categorías, el SPSS dispone del comando ORTHOPLAN.

En la estimación de las utilidades, SPSS utiliza el método ordinario de mínimos cuadrados (OLS). Está demostrado que dicho método resulta tan útil como los demás, incluso cuando se incumplen algunos supuestos paramétricos. Esto hace de CONJOINT una herramienta eficaz y ampliamente extendida. En cuanto a la información que nos proporciona cabe decir que, además de informar de la importancia de los atributos y las utilidades parciales de los niveles de atributos, así como de las correlaciones entre los rangos observados (preferencias) y los rangos predichos por el análisis conjunto, también nos permite especificar tarjetas de reserva (*holdout*) para validar el análisis conjunto o definir tarjetas de simulación para estimar la cuota de mercado o previsión de la demanda.

La base de la que parte esta técnica o método es que los productos o servicios entre los que debe elegir el sujeto en una situación de decisión real son imperfectos y, en consecuencia, debe renunciar a unos atributos en beneficio de otros. Por ellos se define el análisis conjunto como un modelo aditivo. Generalmente, para estudiar estos modelos aditivos, los métodos de investigación tradicionales recurren a preguntar a los sujetos sobre la importancia de cada característica o atributo por separado.

El análisis conjunto a través del perfil completo proporciona un mejor indicador de la importancia relativa de una característica. No olvidemos que los modelos descompositivos son más precisos porque los sujetos toman decisiones analizando los descartes que hay entre las características. Recordemos que estos modelos se basan en que los consumidores toman la decisión de compra considerando simultáneamente todas las características del producto. El consumidor busca un equilibrio entre atributos y el análisis conjunto se utiliza para estudiar estos equilibrios.

El análisis conjunto a través del perfil completo permite desarrollar nuevos servicios o productos, mediante investigaciones con un producto hipotético, con una particular combinación de atributos y niveles, que son preferidos por los consumidores.

ANÁLISIS CONJUNTO Y DISEÑO DE EXPERIMENTOS

El análisis conjunto se basa de modo directo en el diseño de experimentos. De forma muy concreta, se utilizan los diseños factoriales fraccionales y los diseños ortogonales.

También se utilizan los diseños óptimos, ya que en la mayoría de los casos no es conveniente sobrepasar el número de 30 productos a examinar.

En general, un diseño tiende a ser más eficiente cuanto más se acerca a la orthogonalidad y al equilibrio entre factores, de modo que cualquier nivel de un factor esté presente el mismo número de veces en cualquier nivel de otro factor.

Por las razones expresadas, los diseños factoriales fraccionales, los diseños ortogonales y los diseños óptimos son casos particulares de diseños de experimentos muy utilizados en la metodología del análisis conjunto.

SPSS Y EL ANÁLISIS CONJUNTO

ANÁLISIS CONJUNTO A TRAVÉS DE SPSS

Ya sabemos del capítulo anterior que SPSS trata el análisis conjunto siguiendo el denominado *Método del Concepto Completo* o *Perfil Completo (Full Concept)*, denominado así porque considera todos los factores simultáneamente cruzando todos sus niveles entre sí. En este método se le pide al encuestado que elabore un rango, ordenación o puntuación de un prupo de perfiles o tarjetas de acuerdo con sus preferencias.

En cada uno de estos perfiles, se representan todos los factores de interés con una combinación diferente de niveles factoriales (cualidades). De este modo se describe un concepto completo (es decir, un producto o servicio completo) en cada perfil. La tarea del encuestado es elaborar un rango o puntuación de cada perfil desde el menos al más preferido, del menos al más probable de comprar, o de acuerdo a alguna otra escala de preferencia.

A partir de estas clasificaciones (rankings) o puntuaciones, el Análisis Conjunto calcula sólo las puntuaciones de utilidad (utilidades) para cada nivel factorial.

Las puntuaciones de utilidad, análogas a los coeficientes de la regresión, reciben el nombre de *valores parciales (part-worths)* e indican la importancia relativa de cada factor. Tal información puede ser útil cuando se decida qué combinación de niveles factoriales es la mejor para un nuevo producto o servicio y cuando se predigan varios resultados, como las ventas, dadas ciertas combinaciones de los niveles factoriales.

SPSS Y EL MÉTODO DEL CONCEPTO COMPLETO

El *Método del Concepto Completo* se refleja en SPSS mediante las fases siguientes:

Planteamiento del Problema: Las características de un producto se describen en función de sus factores y niveles factoriales. Los factores son los atributos generales del producto, tales como el color, el tamaño, o el precio. En las demás áreas del análisis de datos, estos factores se conocen con el nombre de variables independientes. Los niveles factoriales (también llamados características) son los valores específicos, o categorías, de los factores para un producto particular, tal como azul, grande, y \$10. En las demás áreas del análisis, éstos son los valores de las variables independientes. Para cada caso presentado a los sujetos, se lista un nivel factorial para cada factor en el estudio. El número total de casos que se necesita para representar todas las combinaciones posibles de los niveles factoriales es así igual al de niveles del factor n . El problema de representar tantas combinaciones de niveles factoriales como niveles tenga el factor n es que si los perfiles (o tarjetas) se incluyen para toda posible combinación de los niveles factoriales, hay demasiados perfiles para que los sujetos puedan estimar las puntuaciones. Esta es la razón por la que los estudios del Análisis Conjunto utilizan sólo un pequeño subconjunto de todas las posibles combinaciones, llamado diseño ortogonal.

Generación de un Diseño Ortogonal: Un problema del *Método del Concepto Completo* aparece si están implicados más de dos factores y cada factor tiene más de un par de niveles. El número total de perfiles resultantes de todas las posibles combinaciones de los niveles se hace demasiado grande para que los encuestados las ordenen por rangos o las puntúen de modo significativo. El subconjunto array ortogonal, es un tipo de diseño en el que sólo se consideran los efectos principales y se asume que las interacciones son despreciables. El procedimiento "Generate Orthogonal Design" (Generar Diseño Ortogonal) de SPSS genera un plan de efectos principales, a partir del cual se puede crear un nuevo fichero de datos que contiene el plan ortogonal o reemplazar el fichero de datos activo actual.

Preparación de Tarjetas de Estímulo: La recolección de datos de un diseño de concepto completo requiere que se presenten los estímulos a cada sujeto en un grupo de perfiles individuales. Una vez generado el diseño ortogonal, se debe poner cada ejemplo de un producto completo en una tarjeta (perfil) separada. Esto ayuda al encuestado a enfocarse sólo en el producto realmente bajo evaluación. Se deberían estandarizar también los estímulos, asegurándose de que todos los perfiles son similares en apariencia física, excepto por las diferentes combinaciones de características. El procedimiento "Display Orthogonal Design" (Visualizar Diseño Ortogonal) de SPSS permite imprimir un listado de tarjetas de estímulo, tanto para ser revisadas por el investigador, como para ser presentadas a los encuestados.

Recolección de Datos: Cada encuestado recibe un completo juego de perfiles y se le pide que indique su preferencia para el producto. El investigador puede pedir al encuestado que indique la preferencia asignando una puntuación a cada perfil, donde cuanto mayor sea la puntuación, mayor la preferencia, o asignando un rango a cada perfil (comprendido entre 0 y n donde n es igual al número total de perfiles y un rango inferior significa mayor preferencia), u ordenando los perfiles de los objetos desde el objeto menos al más preferido. Cualquiera que sea el método utilizado, los datos de cada sujeto se codifican en SPSS para ser utilizados por el procedimiento CONJOINT en la estimación de las puntuaciones de utilidad (utilidades).

Estudio de Datos con Análisis Conjunto: El comando CONJOINT es una técnica de análisis categórico que pretende analizar las preferencias asignadas a combinaciones procedentes de estudios conjuntos mediante el *Método del Concepto Completo*. El conjunto de combinaciones para la asignación de preferencias puede estar generado por el procedimiento ORTHOPLAN, o ser introducido directamente por el propio usuario. En los estudios de Análisis Conjunto el investigador asume que se puede definir al producto que se está evaluando en función de unas cuantas características importantes y que cuando un consumidor toma una decisión sobre tal producto, ésta se basa en los balances entre dichas características. Como ninguno de estos productos contendrá todas las mejores características y ninguno concentrará todas las peores, el consumidor decide qué características son importantes y cuáles no. El propósito del Análisis Conjunto es estimar puntuaciones de utilidad, llamadas *valores parciales (part-worths)*, para estas características. Las puntuaciones de utilidad miden la importancia de cada característica en la preferencia global del encuestado.

Interpretación de Resultados: los resultados muestran aspectos importantes, como qué combinación de características es la más preferida, qué niveles particulares influyen más en la preferencia del producto total y la importancia relativa de cada factor. Dado que cada nivel factorial tiene una puntuación de valor parcial, se pueden predecir también los efectos de las combinaciones de los niveles factoriales que no se presentaron realmente en el experimento. La información obtenida a partir de un análisis conjunto puede aplicarse a una amplia variedad de cuestiones de investigación de mercados. Se puede utilizar CONJOINT para investigar áreas como el diseño de productos, el estudio de cuotas de mercado, publicidad estratégica, análisis coste beneficio, y segmentación de mercados. No obstante, CONJOINT es también útil en casi cualquier campo financiero o científico donde sean importantes las mediciones de percepciones o juicios de la gente.

UN EJEMPLO COMPLETO A TRAVÉS DE SPSS

A continuación se desarrolla un ejemplo que ilustra el *Método del Concepto Completo*.

Consideremos una compañía desea lanzar una campaña de mercado para un nuevo limpiador de moquetas y quiere examinar la influencia de los siguientes factores sobre las preferencias del consumidor de artículos de limpieza de moquetas: diseño del paquete (*paquete*), nombre de la marca (*marca*), precio del producto (*precio*), sello de calidad (*sello*) y garantía de devolución del dinero (*dinero*). En el fichero de datos asociado (CPLAN.SAV) se contemplan distintos aspectos de este tipo de artículos. Como niveles factoriales para el diseño del paquete cada uno de los cuales difiere en la localización del cepillo aplacador del producto se consideran A*, B*, C*. Como nombres de marca se consideran *K2R*, *Glory* y *Bisseli*. Como niveles de precios para el producto se consideran \$1.19, \$1.39 y \$1.59 y también se consideran 2 niveles (*Sí* o *No*) para cada uno de los últimos 2 factores (sello y dinero).

Podrían existir otros factores y niveles factoriales que caractericen los limpiadores de moquetas, pero éstos son los únicos en los que está interesada la gerencia de esta compañía de Investigación de Mercados. Por tanto, son los únicos que se considerarán en el análisis conjunto.

De modo análogo a otros diseños experimentales, se podrían querer sólo los factores (las variables independientes) que se cree influirán más en la preferencia o variable respuesta del sujeto (la variable dependiente). Del mismo modo, si un factor tiene más de unos pocos niveles, se podría querer sólo una muestra de niveles realistas que influirán con más probabilidad en la preferencia del producto. Por ejemplo, es improbable que un consumidor tenga una fuerte preferencia por un limpiador sobre otro a causa del precio si los precios difieren poco (por ejemplo, menos de 5 centavos de dólar, en este caso).

Incluso después una selección cuidadosa de los factores y niveles factoriales del estudio, hay todavía demasiados casos para que un sujeto juzgue de un modo significativo. Por ejemplo, el estudio del limpiador de moquetas requeriría 108 casos ($3 \times 3 \times 3 \times 2 \times 2$), que claramente son demasiados para poder presentarlos a un sujeto entrevistado, pues un número razonable de casos normalmente no supera los 30. Afortunadamente, se puede utilizar una alternativa al diseño completamente factorial, llamada *Array Ortogonal*.

Un *Array ortogonal* (una matriz ortogonal de combinaciones) es un subgrupo de todas las posibles combinaciones que todavía permite la estimación de los valores parciales para todos los efectos principales. Se asumen como despreciables las interacciones en las que los valores parciales para un nivel de un factor dependen del nivel de otro factor. Así, en un *Array ortogonal* cada nivel de un factor ocurre con la misma frecuencia que cada nivel de otro factor, es decir, con frecuencias iguales o, al menos, parecidas, asegurando así la independencia de los efectos principales.

Un *Array ortogonal* es la manera óptima de estimar todos los efectos principales. Incluso en el caso de que la estimación mejore a medida que aumenta el número de perfiles, no se pierde realmente información si se omiten algunas combinaciones. Esta es la razón por la que, una vez que se tienen los valores parciales (llamados por el análisis conjunto "utilidades") para cada nivel factorial, se pueden utilizar en las ecuaciones de la predicción para aquellas combinaciones que no fueron evaluadas por los sujetos. Sin embargo, existe una restricción en el número de perfiles que lleva a cabo el análisis conjunto: el número de perfiles debe exceder suficientemente el de factores, para permitir los grados de libertad del error.

Debemos recordar que el Diseño Ortogonal es un modo de ayudar al investigador de mercados a llenar todos y cada uno de los 108 perfiles que corresponderían a todas las combinaciones de las categorías respectivas de las variables que intervienen en el análisis conjunto ($3 \times 3 \times 3 \times 2 \times 2$). Así, en buena lógica, se deberían llenar 108 casos con sus correspondientes columnas, para completar todas las combinaciones posibles. Sin embargo, el Diseño Ortogonal permite evitar este proceso y centrarse sólo en aquellos perfiles interesantes para la compañía vendedora de producto. Por esto, se deberían llenar únicamente aquellos perfiles que le aportan una característica significativa para la Investigación de Mercados que él está realizando. A diferencia de la mayoría de los procedimientos de SPSS, no se requiere un fichero de datos activo antes de generar un diseño ortogonal. Si no se abrió un fichero de datos de trabajo, SPSS crea uno, generando nombres de variables, etiquetas de variables y etiquetas de valores desde las opciones que se seleccionen en los cuadros de diálogo. Si se abrió ya un fichero de datos de trabajo, se le puede reemplazar.

El procedimiento Generar Diseño Ortogonal

El procedimiento *Generate Orthogonal Design* (Generar Diseño Ortogonal) genera un fichero de datos que contiene un diseño o efectos principales, que permite el contraste estadístico de varios factores sin tener que comprobar todas las combinaciones de niveles factoriales. Para poder realizar las operaciones de este tipo es muy importante que SPSS muestre etiquetas de valores en vez de valores de las variables. Por ello hay que comprobar que esté seleccionada la opción *Etiquetas de Valores* en el menú *Ver* (Figura 22-1), o pulsado el botón *Etiquetas de Valores*  en la barra de herramientas estándar del Editor de Datos.

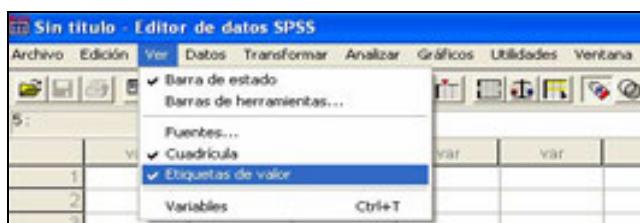


Figura 22-1

A continuación se elige *Datos* → *Diseño ortogonal* → *Generar* (Figura 22-2) para obtener la pantalla de entrada del procedimiento *Generar diseño ortogonal* de la figura 22-3. Introducimos el nombre del primer factor y su etiqueta en la figura 22-4, hacemos clic en el botón *Añadir* y el factor se incorpora al diseño (figura 22-5). Se selecciona con el ratón su nombre sobre la pantalla *Generar diseño ortogonal* (figura 22-6), se hace clic en *Definir valores* y se rellena la pantalla resultante como se indica en la figura 22-7. Se hace clic en *Continuar* y ya aparece la pantalla *Generar diseño* con el nuevo factor y sus valores incorporado (figura 22-8). A continuación se introduce el nombre y la etiqueta de un nuevo factor en la pantalla *Generar diseño* (figura 22-9) y se pulsa *Añadir*. Se selecciona el nuevo factor, se pulsa en *Definir valores* y se rellena la pantalla resultante como se indica en la figura 22-10. Se hace clic en *Continuar* y ya aparece la pantalla *Generar diseño* con los dos factores definidos hasta ahora y sus valores incorporados (figura 22-11). Se repite el proceso hasta generar los 5 factores con sus etiquetas y valores (figura 22-12). Haciendo clic en el botón *Archivo* se puede guardar el diseño con el nombre por defecto (*ortho.sav*) o con cualquier otro a especificar en la figura 11-13.

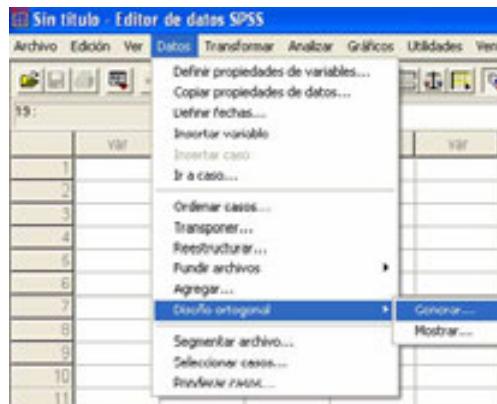


Figura 22-2

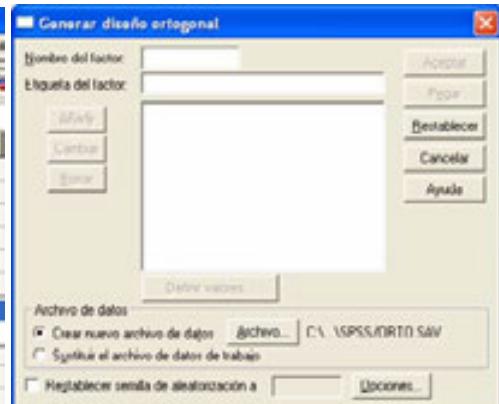


Figura 22-3

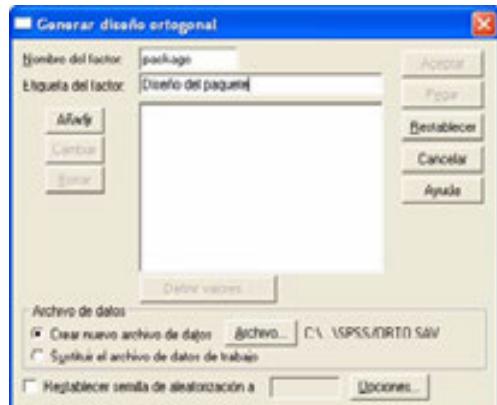


Figura 22-4

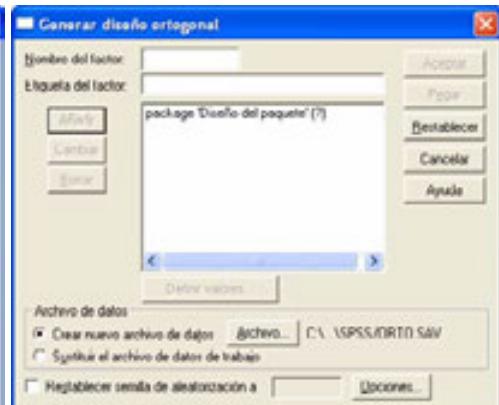


Figura 22-5

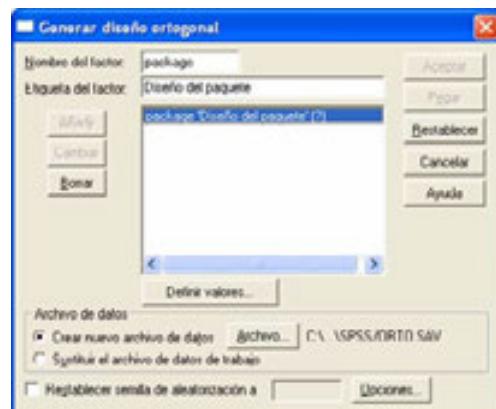


Figura 22-6

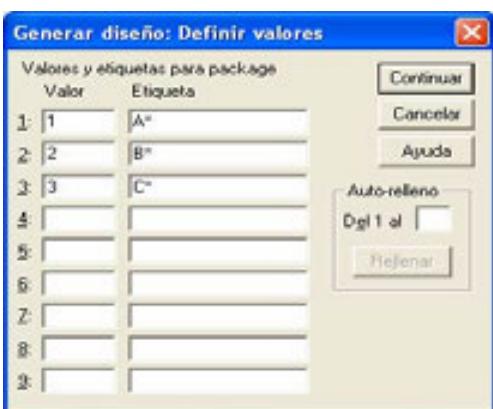


Figura 22-7

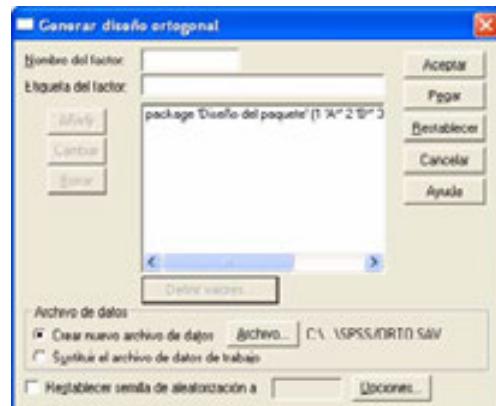


Figura 22-8

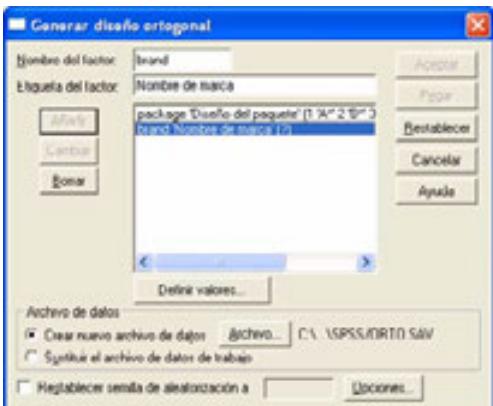


Figura 22-9



Figura 22-10



Figura 22-11

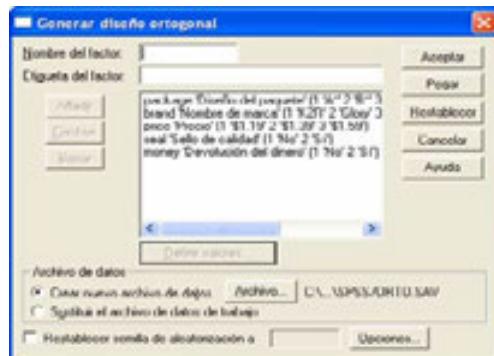


Figura 22-12

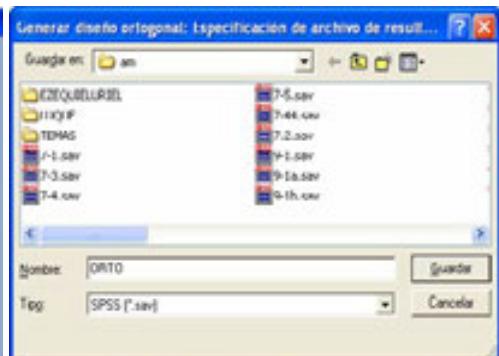


Figura 22-13

Configuración del número de tarjetas de estímulos a generar

En la pantalla *Generar diseño ortogonal* se puede utilizar la casilla *Restablecer semilla de aleatorización* para controlar la generación de los números aleatorios para la creación del diseño ortogonal (figura 22-14). Además, mediante el botón *Opciones* la pantalla *Generar diseño ortogonal* se puede especificar un número mínimo de casos a incluir en el diseño ortogonal y definir el número de casos de reserva prorrteados por los sujetos pero no incluidos por el análisis conjunto (figura 22-15). Los casos de reserva se utilizan en la encuesta, pero el procedimiento Conjoint no los utiliza al estimar las utilidades. Los casos de reserva se generan a partir de otro plan aleatorio, no a partir del plan experimental de efectos principales y no duplican los perfiles experimentales. La opción *Combinar al azar con los otros casos* permite mezclar aleatoriamente los casos de reserva con los casos experimentales. Pulsando Continuar y Aceptar se genera el diseño ortogonal (figura 22-16)

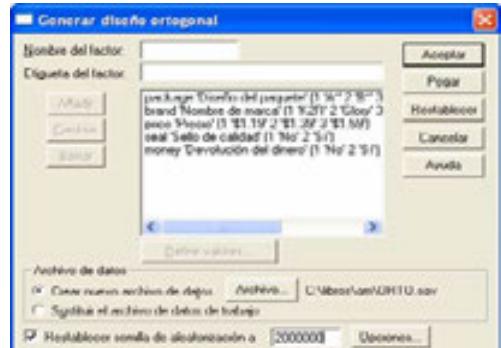


Figura 22-15

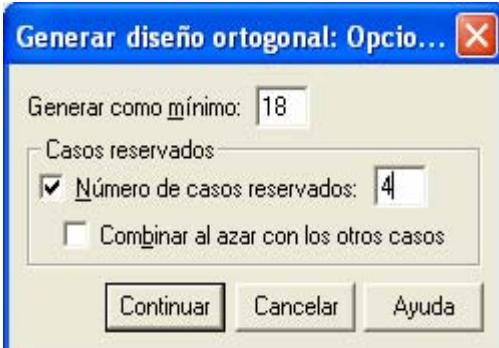


Figura 22-16

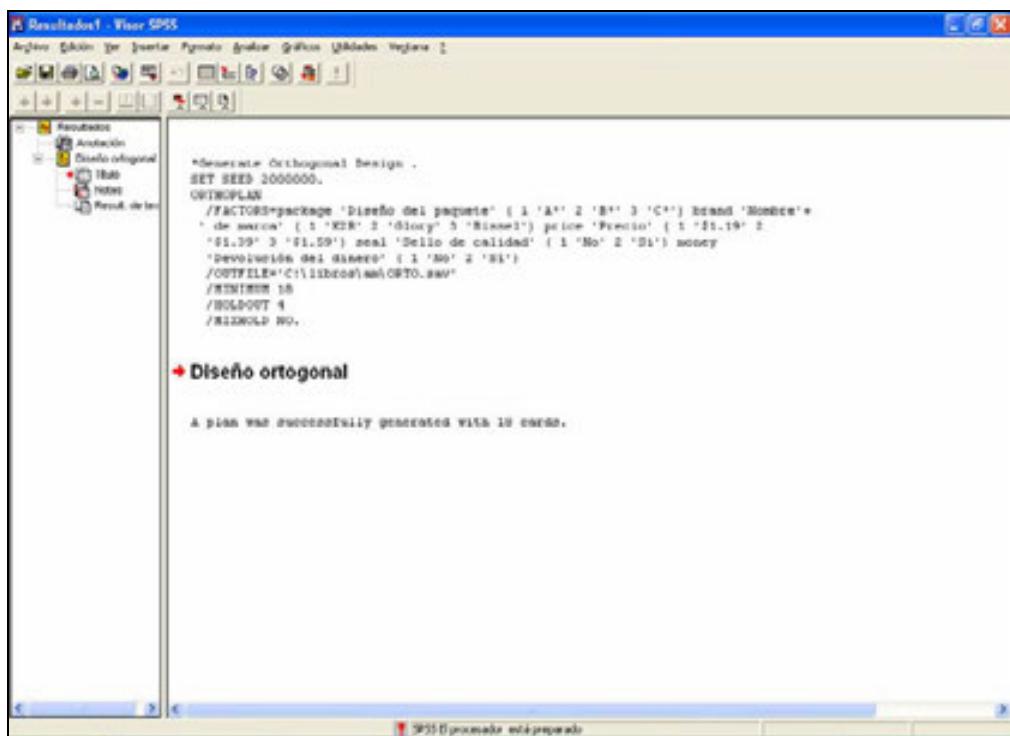


Figura 22-17

Preparación de tarjetas de estímulos

Ya hemos realizado el diseño del plan y ahora debemos situar cada concepto completo en un perfil separado con el objeto de presentárselo a los encuestados en forma de tarjeta (cada caso del diseño ortogonal se muestra como un perfil). Los perfiles pueden visualizarse y personalizarse, siendo posible producir cada concepto como una página separada, añadir títulos y notas a pie de página, controlar el espaciado, etc.

Mediante el procedimiento *Visualizar diseño experimental* es posible mostrar el diseño generado por el procedimiento *Generar diseño ortogonal* (o cualquier otro diseño recogido en un fichero de datos de trabajo) en formato de listado de borrador o como perfiles a mostrar a los sujetos en un análisis conjunto.

Para comenzar cargamos el fichero de datos con el diseño ortogonal ORTO.SAV recién generado (figura 22-18). A continuación elegimos *Datos* → *Diseño ortogonal* → *Mostrar* (Figura 22-19) para obtener la pantalla *Mostrar el diseño* (figura 22-20). La opción *Listado para el experimentador* permite mostar el diseño en formato de borrador diferenciando los perfiles de reserva de los perfiles experimentales y listando los posibles perfiles de simulación de modo separado a continuación de los perfiles experimentales y de reserva.

La opción Perfiles para sujetos produce perfiles que pueden presentarse a los sujetos y la opción Saltos de página después de cada perfil muestra cada perfil en una página nueva. Si se hace clic en el botón *Títulos* de la pantalla *Mostrar el diseño* se puede situar un título y un pié para el perfil (figura 22-21) que aparecerán en el encabezado y en el pié de cada nuevo perfil.

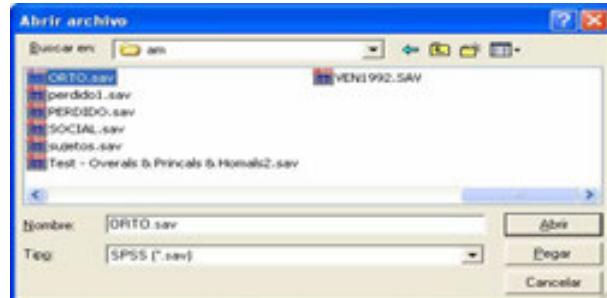


Figura 22-18

Figura 22-19

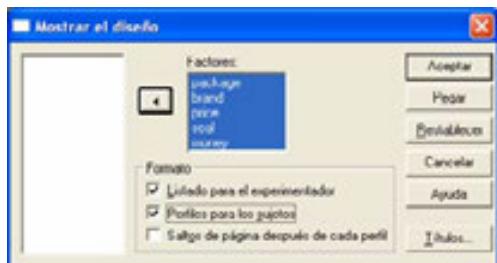


Figura 22-20

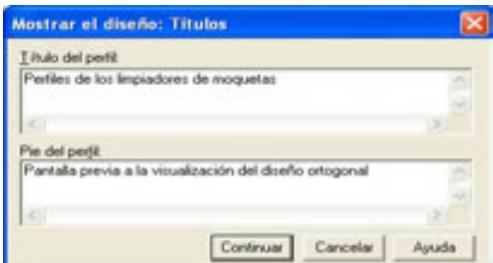


Figura 22-21

Al hacer clic en *Continuar* y en *Aceptar* se muestran las tarjetas del diseño ortogonal generado.

Plancards:

Title: Perfiles de los limpiadores de moquetas

Card 1

Diseño del paquete A*
Nombre de marca Glory
Precio \$1.39
Sello de calidad Sí
Devolución del dinero No

Card 2

Diseño del paquete B*
Nombre de marca K2R
Precio \$1.19
Sello de calidad No
Devolución del dinero No

Card 3

Diseño del paquete B*
Nombre de marca Glory
Precio \$1.39
Sello de calidad No
Devolución del dinero Sí

Card 4

Diseño del paquete C*
Nombre de marca Glory
Precio \$1.59
Sello de calidad No
Devolución del dinero No

Card 5

Diseño del paquete C*
Nombre de marca Bissel
Precio \$1.39
Sello de calidad No
Devolución del dinero No

Card 6

Diseño del paquete A*
Nombre de marca Bissel
Precio \$1.39
Sello de calidad No
Devolución del dinero No

Card 7

Diseño del paquete B*
Nombre de marca Bissel
Precio \$1.59
Sello de calidad Sí
Devolución del dinero No

Card 8

Diseño del paquete A*
Nombre de marca K2R
Precio \$1.59
Sello de calidad No
Devolución del dinero Sí

Card 9

Diseño del paquete C*

-

Nombre de marca K2R
Precio \$1.39
Sello de calidad No
Devolución del dinero No

Card 10
Diseño del paquete C*
Nombre de marca Glory
Precio \$1.19
Sello de calidad No
Devolución del dinero Sí

Card 11
Diseño del paquete C*
Nombre de marca K2R
Precio \$1.59
Sello de calidad Sí
Devolución del dinero No

Card 12
Diseño del paquete B*
Nombre de marca Glory
Precio \$1.59
Sello de calidad No
Devolución del dinero No

Card 13
Diseño del paquete C*
Nombre de marca Bissel
Precio \$1.19
Sello de calidad Sí
Devolución del dinero Sí

Card 14
Diseño del paquete A*
Nombre de marca Glory
Precio \$1.19
Sello de calidad Sí
Devolución del dinero No

Card 15
Diseño del paquete B*
Nombre de marca K2R
Precio \$1.39
Sello de calidad Sí
Devolución del dinero Sí

Card 16
Diseño del paquete A*
Nombre de marca K2R
Precio \$1.19
Sello de calidad No
Devolución del dinero No

Card 17
Diseño del paquete A*
Nombre de marca Bissel
Precio \$1.59
Sello de calidad No
Devolución del dinero Sí

Card 18
—

Diseño del paquete B*
Nombre de marca Bissel
Precio \$1.19
Sello de calidad No

Devolución del dinero No
Card 19 (Holdout)
Diseño del paquete A*
Nombre de marca Bissel
Precio \$1.59
Sello de calidad Sí
Devolución del dinero No
Card 20 (Holdout)
Diseño del paquete C*
Nombre de marca K2R
Precio \$1.19
Sello de calidad Sí
Devolución del dinero No
Card 21 (Holdout)
Diseño del paquete A*
Nombre de marca Glory
Precio \$1.59
Sello de calidad No
Devolución del dinero No
Card 22 (Holdout)
Diseño del paquete A*
Nombre de marca Bissel
Precio \$1.19
Sello de calidad No
Devolución del dinero No

Footer: Pantalla previa a la visualización del diseño ortogonal

—

Perfiles de los limpiadores de moquetas

Diseño del paquete A*
Nombre de marca Glory
Precio \$1.39
Sello de calidad Sí
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete B*
Nombre de marca K2R
Precio \$1.19
Sello de calidad No
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete B*
Nombre de marca Glory
Precio \$1.39
Sello de calidad No
Devolución del dinero Sí

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete C*

Nombre de marca Glory
Precio \$1.59
Sello de calidad No
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete C*
Nombre de marca Bissel
Precio \$1.39
Sello de calidad No
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete A*
Nombre de marca Bissel
Precio \$1.39
Sello de calidad No
Devolución del dinero No

—

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete B*
Nombre de marca Bissel
Precio \$1.59
Sello de calidad Sí
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete A*
Nombre de marca K2R
Precio \$1.59
Sello de calidad No
Devolución del dinero Sí

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete C*
Nombre de marca K2R
Precio \$1.39
Sello de calidad No
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete C*
Nombre de marca Glory
Precio \$1.19
Sello de calidad No
Devolución del dinero Sí

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete C*
Nombre de marca K2R
Precio \$1.59
Sello de calidad Sí
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete B*
Nombre de marca Glory
Precio \$1.59
Sello de calidad No
Devolución del dinero No

—

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete C*
Nombre de marca Bissel
Precio \$1.19
Sello de calidad Sí
Devolución del dinero Sí

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete A*
Nombre de marca Glory
Precio \$1.19
Sello de calidad Sí
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete B*
Nombre de marca K2R
Precio \$1.39
Sello de calidad Sí
Devolución del dinero Sí

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete A*
Nombre de marca K2R
Precio \$1.19
Sello de calidad No
Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal
Perfiles de los limpiadores de moquetas

Diseño del paquete A*

Nombre de marca Bissel

Precio \$1.59

Sello de calidad No

Devolución del dinero Sí

Pantalla previa a la visualización del diseño ortogonal

Perfiles de los limpiadores de moquetas

Diseño del paquete B*

Nombre de marca Bissel

Precio \$1.19

Sello de calidad No

-

Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal

Perfiles de los limpiadores de moquetas

Diseño del paquete A*

Nombre de marca Bissel

Precio \$1.59

Sello de calidad Sí

Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal

Perfiles de los limpiadores de moquetas

Diseño del paquete C*

Nombre de marca K2R

Precio \$1.19

Sello de calidad Sí

Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal

Perfiles de los limpiadores de moquetas

Diseño del paquete A*

Nombre de marca Glory

Precio \$1.59

Sello de calidad No

Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal

Perfiles de los limpiadores de moquetas

Diseño del paquete A*

Nombre de marca Bissel

Precio \$1.19

Sello de calidad No

Devolución del dinero No

Pantalla previa a la visualización del diseño ortogonal

Recogida de los datos

Una vez generado el diseño ortogonal y preparadas las tarjetas de estímulos se abordará la tarea de recoger y analizar los datos. La gran variabilidad de preferencias entre los sujetos puede llevar a tener que seleccionar una muestra de los mismos a partir de la población destino. Este tamaño de muestra suele oscilar entre 100 y 1000 con rango típico entre 300 y 550. En todo caso, el tamaño debe ser tan grande como sea posible para aumentar la fiabilidad de la muestra. Una vez seleccionada la muestra de sujetos, el investigador proporciona el conjunto de tarjetas, o perfiles, a cada encuestado para la recogida de los datos.

Los sujetos pueden registrar los datos asignando una puntuación de preferencia a cada perfil (por ejemplo, se pide a los sujetos que valoren cada perfil asignándole un número de 1 a 100). También se pueden registrar los datos asignando un puesto a cada perfil (un orden según las preferencias del sujeto), es decir, cada perfil obtiene un número entre el 1 y el número total de perfiles los datos. Por último, también pueden ordenarse los perfiles en términos de preferencias registrando el investigador los números de perfiles en el orden dado por cada sujeto.

Para nuestro ejemplo, cuya finalidad es meramente didáctica, se recogen los datos de preferencias de 10 sujetos que ordenan los perfiles del más al menos preferido (cada sujeto asigna un número entre 1 y 22 a cada perfil). El fichero ENCUESTA.SAV recoge los datos (figura 22-22).

	id	pref1	pref2	pref3	pref4	pref5	pref6	pref7	pref8	pref9	pref10	pref11	pref12	pref13	pref14	pref15	pref16	pref17	pref18	pref19	pref20	pref21	pref22
1	1	13	15	1	20	14	7	11	19	3	10	17	11	5	3	6	12	4	21	10	2	22	16
2	2	15	7	2	12	3	11	20	16	21	8	22	8	17	19	1	14	4	9	5	10	13	
3	3	2	18	14	16	22	13	20	10	15	3	1	6	9	5	7	12	19	8	17	21	11	4
4	4	13	10	20	14	2	18	16	22	15	3	1	9	5	6	8	17	11	7	19	4	12	21
5	5	13	18	2	10	20	15	9	5	3	7	11	4	12	22	14	16	1	8	19	21	17	8
6	6	15	2	3	12	18	7	20	10	11	4	9	5	13	16	14	22	8	6	1	21	19	17
7	7	13	7	15	18	2	3	10	20	14	11	19	17	12	1	19	5	4	8	18	16	21	22
8	8	15	7	13	4	6	16	11	22	5	19	21	10	11	3	7	21	14	11	17	14	1	12
9	9	20	9	10	11	4	5	13	15	2	3	12	10	7	1	21	14	16	22	8	6	17	19
10	10	8	21	19	17	4	11	12	7	1	6	9	5	3	15	14	16	22	20	10	13	2	18
11																							

Figura 22-22

Análisis de las preferencias mediante el Análisis Conjunto

Una vez generado el diseño ortogonal (recogido en el fichero ORTO.SAV y recogidos los datos sobre las preferencias en las tarjetas de estímulos provenientes de los sujetos (recogidos en el fichero ENCUESTA.SAV), sólo resta analizar los datos utilizando el procedimiento CONJOINT. Para ejecutar este procedimiento se utilizará la sintaxis de SPSS, abriendo un fichero de sintaxis mediante *Nuevo → Sintaxis* (figura 22-23) y escribiendo la sintaxis de la figura 22-24.

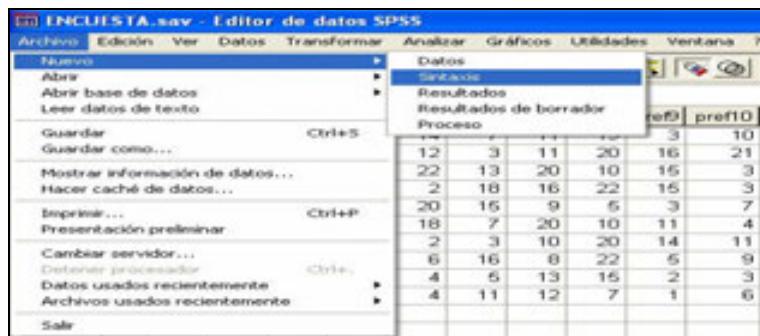


Figura 22-23

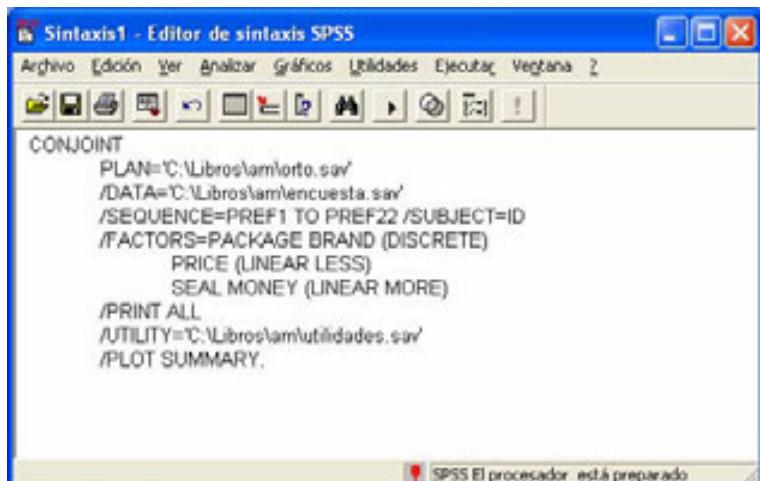


Figura 22-24

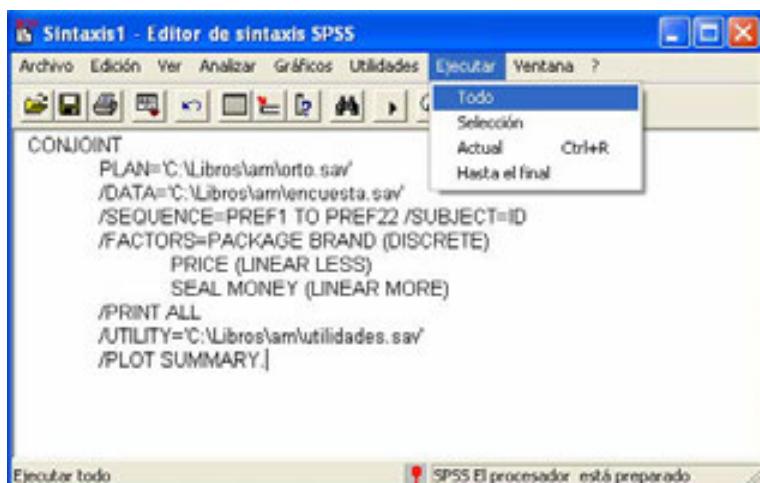


Figura 22-25

La primera línea de la sintaxis es la llamada al procedimiento CONJOINT. El subcomando PLAN identifica el fichero que contiene el diseño ortogonal. El subcomando DATA identifica el fichero que contiene los resultados codificados de la encuesta. El subcomando SEQUENCE indica que los resultados de la encuesta recogidos en el fichero ENCUESTA.SAV han sido codificados en orden secuencial, empezando con la tarjeta más preferida ‘pref1’ y terminando con la menos preferida ‘pref22’, siendo 22 el número de tarjetas generadas. El subcomando SUBJECT identifica la variable que contiene el número del sujeto encuestado. El subcomando FACTORS especifica los factores (variables) definidos en el fichero que contiene el diseño ortogonal identificado por el subcomando PLAN. Se observa que los factores *package* y *brand* se definen como discretos (variables categóricas) y no se hace ninguna asunción sobre la relación entre los niveles y los datos. El factor *price* se define como menos lineal (variable lineal para la que los consumidores prefieren los precios más bajos). Los factores *seal* y *money* se definen como más lineales (variables lineales para las que se supone que los consumidores prefieren aquella para la que el producto tenga sello de calidad y se garantice la devolución del dinero).

El subcomando PRINT permite controlar la salidas de texto y ALL especifica que se presenten tanto los resultados de los datos experimentales, como los de simulación. El subcomando UTILITY identifica el fichero en el que CONJOINT guardará las utilidades calculadas generándose un caso por cada sujeto encuestado. El subcomando PLOT solicita las salidas gráficas. La palabra clave SUMMARY produce un diagrama de barras par cada variable, mostrando las puntuaciones de la utilidad para cada categoría de esa variable y un gráfico que muestra las puntuaciones de importancia de resumen por sujetos con la palabra clave SUBJECT. Con *Ejecutar → Todo* (figura 22-25) se tiene la salida del procedimiento CONJOINT, que empieza con los factores del diseño ortogonal (figura 22-26).

The screenshot shows the SPSS Results window titled "Resultados2 - Visor SPSS". The menu bar includes Archivo, Edición, Ver, Insertar, Formato, Análisis, Gráficos, Utilidades, Vistas, etc. The toolbar has various icons for file operations. The left pane displays a tree view of results, with "Análisis conjunto" expanded, showing sub-items like "Título", "Notas", "Result. de tex...", and several "Resumen de..." entries. The right pane contains two main sections: "CONJOINT" and "Análisis conjunto".

```

CONJOINT
/PLAN='C:\LIBROS\ANALITO\ANALITO.SAV'
/DATA='C:\LIBROS\ANALITO\ENCUESTA.SAV'
/SEQUENCE=PREF1 TO PREF22 /SUBJECT=ID
/FACTORS=PACKAGE BRAND (DISCRETE)
PRICE (LINEAR LESS)
SEAL MONEY (LINEAR MORE)
/PRINT ALL
/UTILITY='C:\LIBROS\ANALITO\UTILIDADES.SAV'
/PLOT SUMMARY.

```

Análisis conjunto

Factor	Model	Levels	Label
PACKAGE	d	3	Diseño del paquete
BRAND	d	3	Nombre de marca
PRICE	i<	3	Precio
SEAL	i>	2	Sello de calidad
MONEY	i>	2	Devolución del dinero

(Modelos: d-discreto, i-lineal, i-ideal, ai-antídeval, <-less, >-more)

All the factors are orthogonal.

Figura 22-26

Interpretación de las salidas del Análisis Conjunto

Los resultados del procedimiento Conjoint se ofrecen ordenadamente por sujetos. A continuación se muestra la salida para el primer sujeto

SUBJECT NAME:	1	
Importance	Utility(s.e.)	Factor ** Reversed (1 reversal)
7,21		PACKAGE Diseño del paquete
	,0000(,6303)	A*
	-,6667(,6303)	B*
	,6667(,6303)	C*
12,61		BRAND Nombre de marca
	-1,3333(,6303)	K2R
	1,0000(,6303)	Glory
	,3333(,6303)	Bissel
4,50		PRICE ** Precio
	,4167(,5458)	\$1.19
	,8333(1,0916)	\$1.39
	1,2500(1,6375)	\$1.59
	B = ,4167(,5458)	
48,65		SEAL Sello de calidad
	9,0000(,9454)	-- No
	18,0000(1,8908)	---- Sí
	B = 9,0000(,9454)	
27,03		MONEY Devolución del dinero
	5,0000(,9454)	- No
	10,0000(1,8908)	-- Sí
	B = 5,0000(,9454)	
	-10,000(2,1373)	CONSTANT
Pearson's R	= ,962	Significance = ,0000
Kendall's tau	= ,869	Significance = ,0000
Kendall's tau	= ,667 for 4 holdouts	Significance = ,0871

En la salida se observan las puntuaciones de la utilidad y su error estándar para cada nivel factorial. La utilidad total de una combinación específica se halla sumando los valores de sus puntuaciones correspondientes. Por ejemplo, la utilidad total de un limpiador con diseño de paquete (*package*) B*, marca (*brand*) K2R, precio (*price*) \$1,19, no sellado (*no seal*) y sin garantía de devolución del dinero (*no money back*) sería:

```
utility(package B*) + utility(K2R) + utility($1.19) +
utility(no seal) + utility(no money-back) + constante
```

Realizando la valoración de la utilidad total del limpiador anterior tenemos:

$$(-0,6667) + (-1,3333) + 2,4792 + 9,0000 + 5,0000 + (-12,0620) = 2,4172$$

Las utilidades totales distan algo de los datos observados (aunque teóricamente deberían de coincidir). El error estándar de cada utilidad indica el grado de ajuste del modelo a los datos del sujeto particular considerado. En la salida para el sujeto 1 se observan errores estándar altos para *price*, con lo que puede ser que el modelo lineal no sea el más adecuado para este factor en el caso de este sujeto.

La columna más a la izquierda de la salida anterior presenta las puntuaciones de la importancia de cada factor, junto con un gráfico de barras para dar una idea de cómo se comparan los factores. Las puntuaciones de la importancia se calculan tomando el rango de la utilidad para el factor particular y dividiéndolo por la suma de todos los rangos de las utilidades.

Los estadísticos R de Pearson y Tau de Kendall indican también el grado de ajuste de los datos al modelo y representan las correlaciones entre las preferencias observadas y estimadas y, por tanto, deberían ser siempre muy altas.

La salida de Conjoint ofrece este mismo análisis para los 10 sujetos.

Cuando se utiliza el subcomando SUBJECT con Conjoint, se consiguen, además de los resultados para cada sujeto, unos resultados medios para todo el grupo denominados resultados agrupados del comando Conjoint y etiquetados SUBFILE SUMMARY *Averaged Importante* (resumen del subfichero en importancia media) y que se presentarán a continuación.

Al final de la salida se observa el resumen de reversiones y de simulaciones, que ofrece las probabilidades de elegir los perfiles de simulación particulares como perfiles más preferidos, bajo el modelos de probabilidad de elección de la *Máxima Utilidad* (probabilidad de elegir un perfil como el más preferido), bajo el modelo BTL (*Bradley-Terry-Luce*) que calcula la probabilidad de elegir un perfil como el más preferido dividiendo la utilidad del perfil entre la suma de todas las utilidades totales de la simulación, y bajo el modelo Logit, que es similar la modelo BTL, pero que utiliza el logaritmo de las utilidades en vez de las utilidades mismas.

Se observa que para los diez sujetos de estudio, los tres modelos indican que el perfil de simulación 2 sería el más preferido.

SUBFILE SUMMARY

Averaged Importance	Utility	Factor	
35,63	-2,2333 1,8667 ,3667	PACKAGE -- -- -- --	Diseño del paquete A* B* C*
14,91	,3667 -,3500 -,0167	BRAND -- --	Nombre de marca K2R Glory Bissel
29,41	-1,1083 -2,2167 -3,3250 B = -1,1083	PRICE - -- -- --- --- ---	Precio \$1.19 \$1.39 \$1.59
11,17	2,0000 4,0000 B = 2,0000	SEAL -- ----	Sello de calidad No Sí
8,87	1,2500 2,5000 B = 1,2500	MONEY - --- --- ---	Devolución del dinero No Sí
	7,3833	CONSTANT	
Pearson's R	= ,982		Significance = ,0000
Kendall's tau	= ,892		Significance = ,0000
Kendall's tau	= ,667 for 4 holdouts		Significance = ,0871

SUBFILE SUMMARY

Reversal Summary:

2 subjects had 2 reversals
 3 subjects had 1 reversals

Reversals by factor:

PRICE	3
MONEY	2
SEAL	2
BRAND	0
PACKAGE	0

Reversal index:

Page	Reversals	Subject
1	1	1
2	2	2
3	0	3
4	0	4
5	0	5
6	1	6
7	0	7
8	0	8
9	1	9
10	2	10

El subcomando PLOT de Conjoint aporta un modo gráfico de observar los resultados del grupo. La palabra clave SUMMARY produce un diagrama de barras para cada variable mostrando las puntuaciones de la utilidad para cada categoría de esa variable (figuras 22-27 a 22-31) y un gráfico que muestra las puntuaciones de importancia del resumen por sujetos con la palabra clave SUBJECT (figura 22-32).

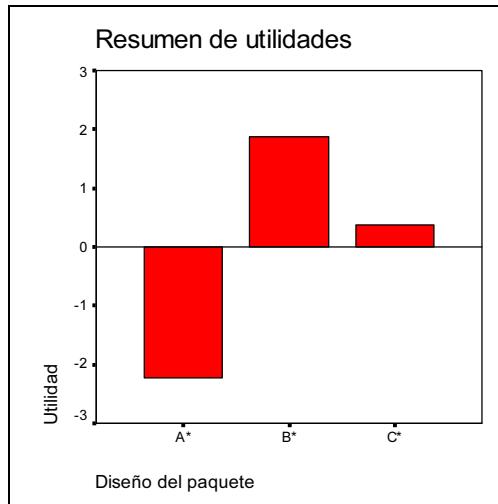


Figura 22-27

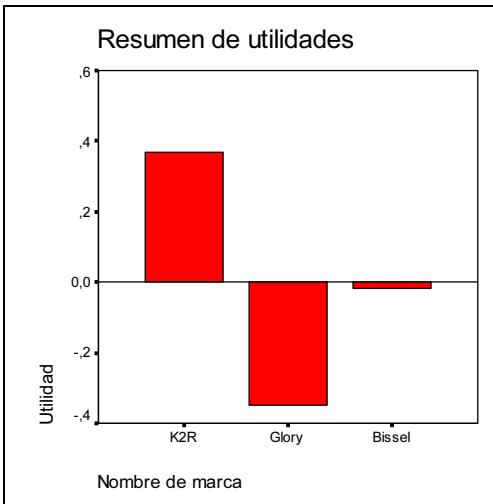


Figura 22-28

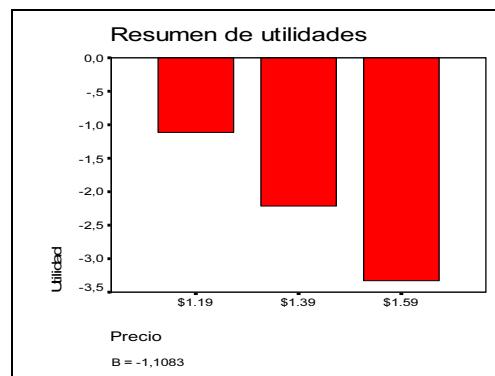


Figura 22-29

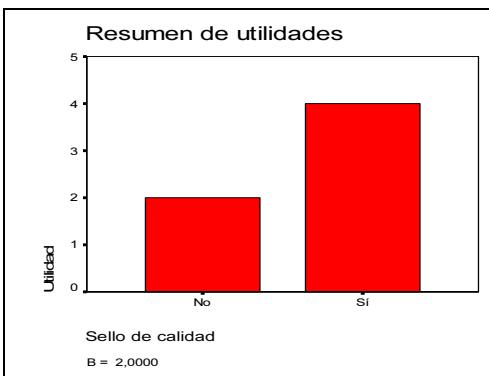


Figura 22-30

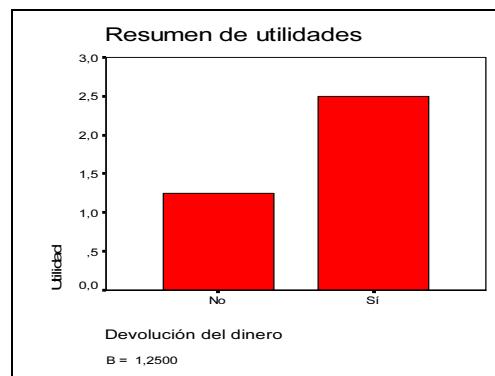


Figura 22-31

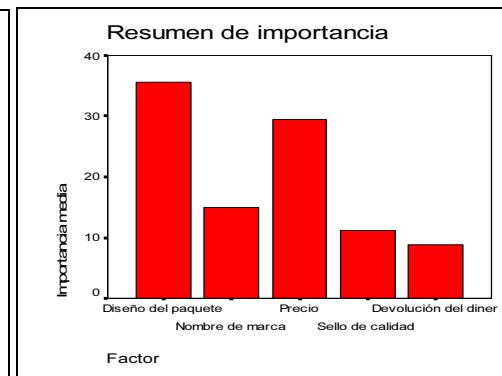


Figura 22-32

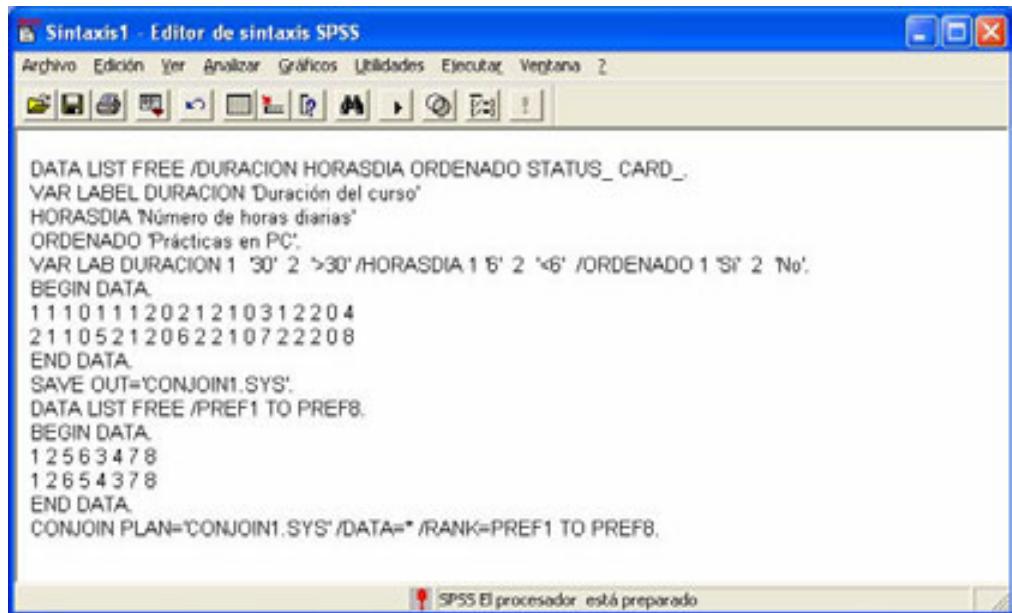
El subcomando UTILITY crea un fichero de datos de SPSS (figura 22-33) que contiene para cada sujeto las utilidades para los factores DISCRETE, la pendiente y las funciones cuadráticas para los factores LINEAL, DEAL y ANTIDEAL (etiquetas B y C en la salida), la constante de regresión y las puntuaciones estimadas de las preferencias. Estos valores se pueden utilizar en análisis posteriores para realizar gráficos adicionales y gráficos con otros procedimientos.

	id	constant	packag1	packag2	packag3	brand1	brand2	brand3	price_1	seal_1	money_1	score1	score2
1	1.00	-10.00	.00	-.87	.87	-1.33	1.00	.33	.42	2.00	5.00	14.83	2.42
2	2.00	7.83	1.00	6.00	5.00	2.00	1.67	.33	1.17	1.00	1.50	8.00	16.50
3	3.00	20.60	.00	4.00	2.00	.33	.17	.17	6.00	.75	.00	9.83	17.58
4	4.00	15.50	.00	-.17	.17	.33	.00	.33	6.00	1.50	3.00	9.50	14.17
5	5.00	13.50	6.00	3.00	3.00	-.17	1.00	.83	3.33	1.00	1.00	2.83	16.33
6	6.00	10.67	6.00	6.00	.00	1.33	.50	1.83	.56	.25	.25	3.75	17.42
7	7.00	1.33	-4.00	4.33	-.33	-1.33	-.33	1.67	-1.75	5.50	3.25	7.75	11.33
8	8.00	6.33	-1.33	.17	1.17	1.00	-3.17	2.17	-.42	1.25	1.75	5.25	10.00
9	9.00	10.33	-6.00	.00	6.00	1.17	.83	-2.00	-.56	.50	-.25	4.75	11.17
10	10.00	-2.17	2.00	-2.00	.00	.50	-.50	8.00	-.25	.00	11.00	1.50	

Figura 22-33

Ejercicio 22-1. Se trata de estudiar las preferencias sobre un determinado curso de formación. Se suponen tres factores como importantes para determinar las citadas preferencias: la duración total del curso (variable DURACIÓN), el número de horas diarias de clase (variable HORASDÍA) y la existencia o no de prácticas en PC (variable ORDENADO). A la variable DURACIÓN se le asigna el valor 0 si el curso dura menos de 30 horas, el valor 1 si el curso dura 30 horas y el valor 2 si el curso dura más de 30 horas. A la variable HORASDIA se le asigna el valor 0 si se imparten menos de 6 horas diarias, el valor 1 si se imparten 6 horas diarias y el valor 2 si se imparten menos de 6 horas diarias. A la variable ORDENADO se le asigna el valor 1 si se realizan prácticas en PC y el valor 2 en caso contrario. Se considera el diseño factorial completo de las 8 combinaciones de niveles de los tres factores y se recogen datos de 2 sujetos que ordenan las preferencias de la forma 1 2 5 6 3 4 7 8 y 1 2 6 5 4 3 7 8 mediante los rangos PREF1 y PREF2. Obtener las estimaciones de las utilidades de cada uno de los niveles factoriales y los coeficientes de correlación de Pearson y Tau de Kendall entre las utilidades estimadas y observadas para el conjunto de los dos sujetos.

Como el diseño a utilizar es factorial completo con sólo tres factores con pocos niveles cada uno y la recogida de datos está referida sólo a dos sujetos, utilizaremos un fichero de sintaxis SPSS abriendo un fichero de sintaxis mediante Nuevo → Sintaxis y escribimos el código de la figura 22-33. Al elegir Ejecutar → Todo se obtiene la salida del procedimiento Conjoint (figura 22-34).



```

Sintaxis1 - Editor de sintaxis SPSS
Archivo Edición Ver Analizar Gráficos Utilidades Ejecutar Ventana ?
DATA LIST FREE /DURACION HORASDIA ORDENADO STATUS_CARD_
VAR LABEL DURACION 'Duración del curso'
HORASDIA 'Número de horas diarias'
ORDENADO 'Prácticas en PC'
VAR LAB DURACION 1 '30' 2 '>30'/HORASDIA 1 '6' 2 '<6' /ORDENADO 1 'Sí' 2 'No'.
BEGIN DATA.
11101112021210312204
21105212062210722208
END DATA.
SAVE OUT='CONJOIN1.SYS'.
DATA LIST FREE /PREF1 TO PREF8.
BEGIN DATA.
12563478
12654378
END DATA.
CONJOIN PLAN='CONJOIN1.SYS' /DATA=* /RANK=PREF1 TO PREF8.

SPSS El procesador está preparado

```

Figura 22-33

The screenshot shows the SPSS Results window with the title "Resultados - Vista SPSS". The menu bar includes Archivo, Edición, Ver, Insertar, Formato, Analizar, Gráficos, Métodos, Vigneta, and Ayuda. The toolbar has icons for various operations like Open, Save, Print, and Paste. The left pane displays a tree structure with "Resultados" expanded, showing "Análisis", "Análisis conjunto", "Título", "Notas", and "Resumen de los resultados". The main pane contains the following text:

```

DATA LIST FREE /DURACION MORASDIA ORDENADO STATUS_ CARD_.
VAR LABEL DURACION 'Duración del cuchillo'
MORASDIA 'Número de horas diarias'
ORDENADO 'Prácticas en PC'.
VAR LAB DURACION 1 '30' 2 '>30' /MORASDIA 1 '6' 2 '<6' /ORDENADO 1 'SI' 2
  'NO'.
BEGIN DATA.
1 1 0 1 1 1 2 0 2 1 2 1 0 3 1 3 2 0 4
2 1 1 0 5 2 1 2 0 4 2 2 1 0 7 2 2 2 0 8
END DATA.
SAVE OUT='CONJOINT1.SPS'.
DATA LIST FREE /PREF1 TO PREF9.
BEGIN DATA.
1 2 5 6 3 4 7 8
1 2 6 5 4 3 7 8
END DATA.
COMPOIN PLAN='CONJOINT1.SPS' /DATA=* /RANK=PREF1 TO PREF9.

```

Análisis conjunto

```

Factor Model Levels Label
DURACION d 2 1 '30' 2 '>30'
MORASDIA d 2 1 '6' 2 '<6'
ORDENADO d 2 1 'SI' 2 'No'
(Modelos: d=dimension, l=linear, i=ideal, a=antiideal, c=less, >more)

All the factors are orthogonal.
□

Averiadas  Representante Utilidad  Preferencias

```

SPSS El procesador está preparado

Figura 22-34

La salida completa muestra unos estadísticos de Pearson y Kendal muy cercanos a la unidad, lo que indica que el grado de ajuste de los datos al modelo es muy alto. Además la significación de esos coeficientes es muy elevada y las correlaciones entre las preferencias observadas y estimadas son muy altas.

No olvidemos que las utilidades observadas y estimadas para cada uno de los dos individuos debieran de coincidir teóricamente, por lo que mientras más altos sean los distintos coeficientes de correlación entre ambas utilidades, mejor será el ajuste.

Se obtienen también las estimaciones de cada uno de los distintos niveles factoriales.

La salida completa se presenta a continuación:

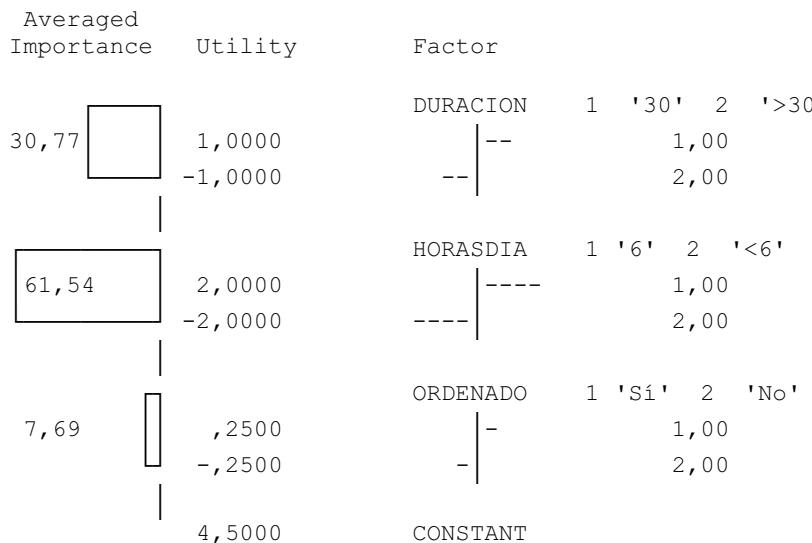
```

Factor Model Levels Label
DURACION d 2 1 '30' 2 '>30'
HORASDIA d 2 1 '6' 2 '<6'
ORDENADO d 2 1 'Sí' 2 'No'
(Models: d=discrete, l=linear, i=ideal, ai=antiideal, <=less, >=more)

```

All the factors are orthogonal.

-



Pearson's R = ,994 Significance = ,0000

Kendall's tau = ,964 Significance = ,0005

-

SUBFILE SUMMARY

No reversals occurred in this split file group.

Ejercicio 22-2. Generar el diseño ortogonal asociado al problema anterior.

Comenzamos generando el diseño ortogonal mediante *Datos → Diseño ortogonal → Generar* (Figura 22-35) para obtener la pantalla de entrada del procedimiento *Generar diseño ortogonal* de la figura 22-36. Introducimos el nombre del primer factor y su etiqueta en la figura 22-37, hacemos clic en el botón *Añadir* y el factor se incorpora al diseño (figura 22-38). Se selecciona con el ratón su nombre sobre la pantalla *Generar diseño ortogonal* (figura 22-39), se hace clic en *Definir valores* y se rellena la pantalla resultante como se indica en la figura 22-40.

Se hace clic en *Continuar* y ya aparece la pantalla *Generar diseño* con el nuevo factor y sus valores incorporado (figura 22-41). A continuación se introduce el nombre y la etiqueta de un nuevo factor en la pantalla *Generar diseño* (figura 22-42) y se pulsa *Añadir*. Se selecciona el nuevo factor, se pulsa en *Definir valores* y se rellena la pantalla resultante como se indica en la figura 22-43. Se hace clic en *Continuar* y ya aparece la pantalla *Generar diseño* con los dos factores definidos hasta ahora y sus valores incorporados (figura 22-44). Se repite el proceso con el factor que falta (figuras 22-45 a 22-47). Haciendo clic en el botón *Archivo* se puede guardar el diseño ortogonal con el nombre *orto1.sav* (figura 22-48). Al hacer clic en *Aceptar* en la figura 22-47 se obtiene el diseño ortogonal con 9 tarjetas de estímulo (figuras 22-49 y 22-50).

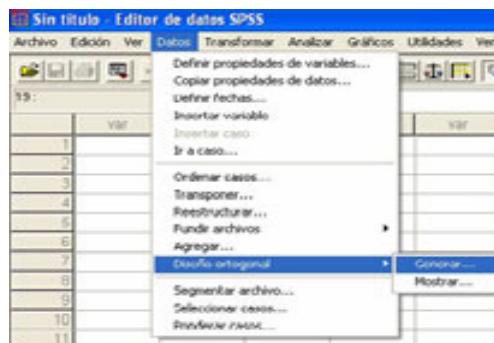


Figura 22-35

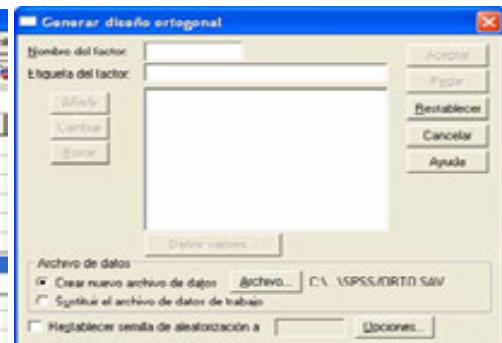


Figura 22-36



Figura 22-37

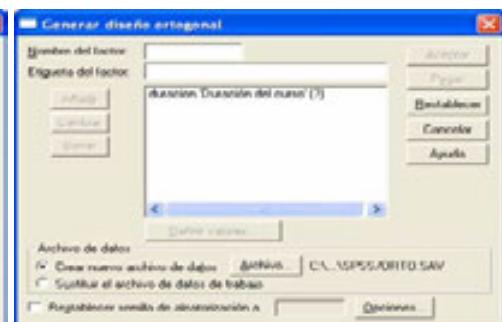


Figura 22-38

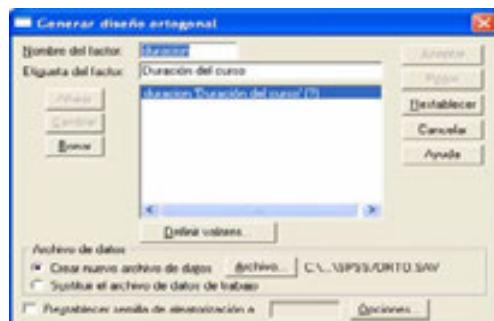


Figura 22-39

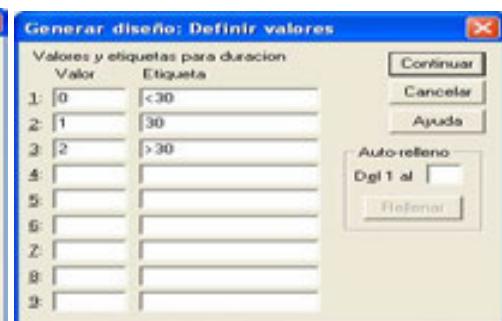


Figura 22-40

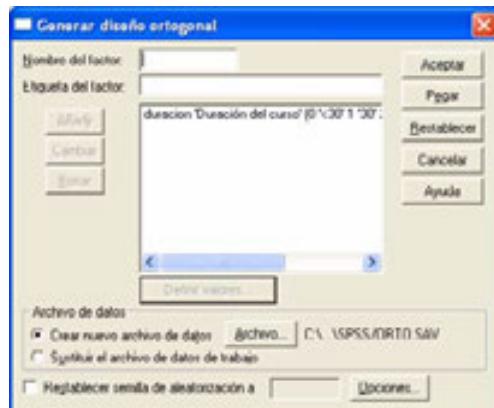


Figura 22-41

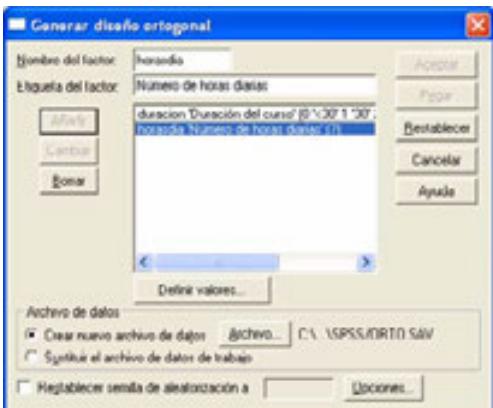


Figura 22-42



Figura 22-43

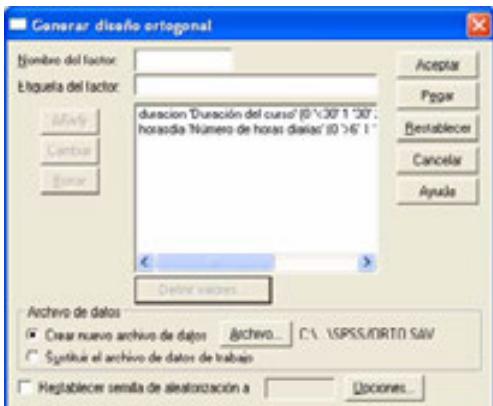


Figura 22-44

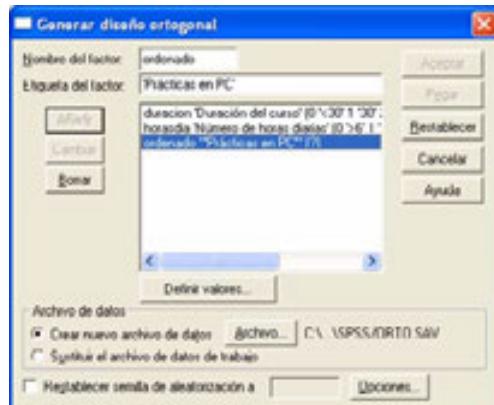


Figura 22-45

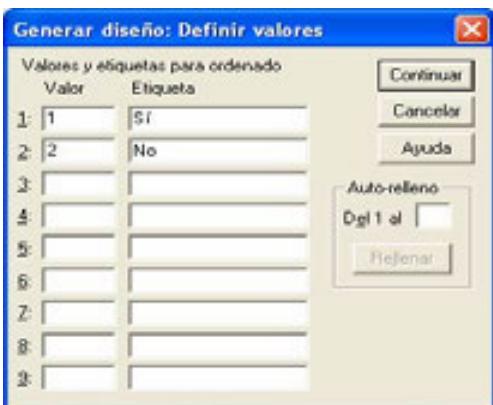


Figura 22-46

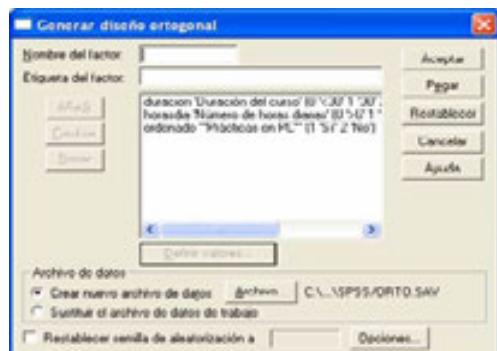


Figura 22-47

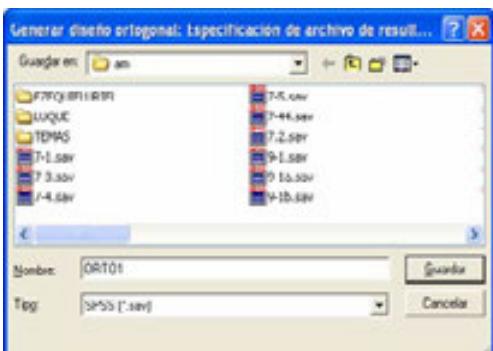


Figura 22-48

Resumen
Diseño ortogonal
Título
Notas
Resultados

```
*Generate Orthogonal Design .
ORTOPLAN
/FACTORS=duración "Duración del curso" ( 0 '<30' 1 '30' 2 '>30') horasdiá
"Número de horas diarias" ( 0 '>6' 1 '6' 2 '<6') ordenado "Prácticas en"+
= PC+" ( 1 'Sí' 2 'No')
/OUTFILE="C:\libros\am\ORTO1.sav" .
```

◆ Diseño ortogonal

A plan was successfully generated with 9 cards.

Figura 22-49

	duración	horasdiá	ordenado	status_	card_
1	>30	<6	Si	Diseño	1
2	<30	6	Si	Diseño	2
3	>30	>6	Si	Diseño	3
4	<30	<6	No	Diseño	4
5	30	<6	Si	Diseño	5
6	>30	6	No	Diseño	6
7	30	6	Si	Diseño	7
8	30	>6	No	Diseño	8
9	<30	>6	Si	Diseño	9

Figura 22-50

ÍNDICE ALFABÉTICO

A

- ACA (Adaptative Conjoint Análisis).....* 599
Adecuación individual 176
Adecuación muestral global al modelo factorial.... 176
Agrupación de centroides..... 442
Agrupación de medianas 442
AID (Automatie Interaction Detector Method).... 433
*Ajuste y preferencias en el escalamiento
multidimensional.....* 285
Algoritmo ALSCAL 286
Algoritmo de las H-medias..... 425
Algoritmo de las K-medias 425
Algoritmo INDSCAL..... 296
Amplitud intercuartil..... 34, 69
Análisis ANOVA de la varianza..... 495
Análisis canónico (correlación canónica)..... 5
*Análisis cluster... 1, 14, 15, 208, 284, 417, 418, 419,
424, 435, 437, 439, 440, 441, 444, 447, 448, 452*
Análisis conjunto..... 9, 595
Análisis conjunto a través de SPSS..... 605
Análisis conjunto a través del perfil completo 603
*Análisis conjunto como una técnica
multivariante de la dependencia.....* 597
Análisis conjunto y diseño de experimentos .. 604
Análisis de componentes principales categóricas ... 280,
281, 304
*Análisis de componentes principales categóricas
(CATPCA)* 276, 303
*Análisis de componentes principales categóricas
con SPSS.....* 281, 304, 305
Análisis de componentes principales estándar ... 280,
281, 304
Análisis de conglomerados (análisis cluster).14, 417

- Análisis de conglomerados en dos fases ...* 433, 445, 452
Análisis de conglomerados jerárquicos.... 437, 438, 441
Análisis de Correlación Canónica 278, 279
*Análisis de correlación canónica categórico...*279**
Análisis de correlación canónica no lineal 279,
303, 317, 319
*Análisis de correlación canónica no lineal
(OVERALS)* 303
Análisis de correspondencias .2, 11, 13, 252, 268, 303
Análisis de correspondencias múltiple.. 13, 236, 246,
260, 308
Análisis de correspondencias múltiples .. 13, 303
Análisis de correspondencias múltiples HOMALS.. 279
*Análisis de correspondencias múltiples u
homogeneidades (HOMALS).....* 276
Análisis de correspondencias simple.. 13, 236, 237, 246
Análisis de Correspondencias Simples (ANACOR)... 276
Análisis de homogeneidades..... 260, 266, 274
Análisis de la covarianza simple..... 7, 514
Análisis de la covarianza múltiple..... 8, 516
Análisis de la dispersión 411
Análisis de la supervivencia..... 559, 560, 562
Análisis de la varianza con varios factores 503
Análisis de la varianza múltiple..... 8, 515
Análisis de la varianza múltiple MANOVA. 8, 515, 516
Análisis de la varianza simple..... 7
Análisis de la varianza simple (ANOVA) 496
Análisis de la varianza simple (un solo factor) 497
*Análisis de las preferencias mediante el Análisis
Conjunto* 621
Análisis de los datos ausentes 21, 39
Análisis de los residuos 68, 378, 379

<i>Análisis de los residuos de los parámetros del modelo</i>	378
<i>Análisis de los residuos en las celdas</i>	379
<i>Análisis de los residuos tipificados</i>	380
<i>Análisis de mínimos cuadrados ponderados (MCP)</i>	527
<i>Análisis de Regresión Múltiple</i>	279
<i>Análisis de supervivencia de Kaplan-Meier</i>	562, 580, 583
<i>Análisis de supervivencia de Kaplan-Meier con SPSS</i>	580
<i>Análisis de varianza para variables dependientes múltiples</i>	527
<i>Análisis del grado de asociación entre variables cualitativas</i>	359
<i>Análisis discriminante</i>	457
<i>Análisis discriminante</i> 2, 5, 279, 280, 457, 483, 487	
<i>Análisis discriminante canónico</i>	458, 473
<i>Análisis discriminante múltiple</i>	470, 472
<i>Análisis EMD mediante ALSCAL</i>	324
<i>Análisis en componentes principales</i>	11, 138
<i>Analisis en R^a</i>	227
<i>Análisis en R^p</i>	223, 231, 234, 241
<i>Analisis exploratorio de los datos con SPSS</i>	69
<i>Análisis exploratorio y gráfico de los datos</i>	21
<i>Análisis exploratorio y gráfico de los datos</i>	22
<i>Análisis factorial</i>	12, 155
<i>Análisis factorial confirmatorio</i>	192
<i>Análisis factorial de correlaciones</i>	173
<i>Análisis factorial de correspondencias</i>	237, 238, 240, 241, 243, 246
<i>Análisis factorial de correspondencias</i>	236
<i>Análisis factorial exploratorio</i>	192
<i>Análisis factorial exploratorio y confirmatorio</i>	191
<i>Análisis general de los métodos factoriales</i>	222
<i>Análisis logit</i>	388
<i>Análisis loglineal general</i>	388, 397
<i>Análisis modal de Wishart</i>	424
<i>Análisis multivariante de la dependencia</i>	10
<i>Análisis multivariante de la interdependencia</i>	15
<i>Análisis no lineal de componentes principales (CATPCA)</i>	281, 304
<i>Análisis no níquel de correlación canónica (OVERALS)</i>	276
<i>Análisis previo de los datos</i>	21
<i>Análisis probit</i>	574, 575
<i>Análisis y detección de valores atípicos</i>	48
<i>Analizar la autocorrelación de un modelo</i>	66
<i>Aplicaciones del análisis conjunto</i>	600
<i>Aplicaciones del análisis de correspondencias</i>	266
<i>Aplicaciones del MDS</i>	301
<i>Aproximación de casos completos</i>	46, 89, 117
<i>Aproximación de perfil completo</i>	603
<i>Array ortogonal</i>	608, 609
<i>Asociación en modelos logarítmico lineales</i>	373
<i>Asociación entre las variables independientes y la dependiente</i>	411
<i>Asociación entre variables cualitativas</i>	359
<i>Asociación parcial</i>	373
<i>Autocorrelación</i>	65
<i>Autocorrelación o correlación serial</i>	55, 65, 68
<i>Axioma de desigualdad triangular</i>	293
<i>Axioma de no-negatividad</i>	293
<i>Axioma de simetría</i>	293
<i>Ayudas a la interpretación para cada fila y columna</i>	249
C	
<i>Cantidad de información</i>	122, 219, 220, 221, 223, 226, 228, 229, 234
<i>Cantidad de información de la nube de puntos</i> ...	223
<i>Cantidad de información y distancias</i>	219
<i>Carga factorial o saturación</i>	158
<i>Cargas factoriales</i> 144, 145, 146, 152, 153, 160, 164, 170, 181, 182, 183, 184, 185, 186, 190, 192, 207, 216, 217	
<i>Cargas factoriales estimadas</i>	160
<i>Censura de intervalo</i>	560
<i>Censura por la derecha</i>	559
<i>Censura por la izquierda</i>	559
<i>Centro de gravedad</i> ..	139, 141, 142, 165, 166, 219, 220, 221, 222, 223, 232, 234, 242, 244, 248, 249, 420
<i>Centro de gravedad de la nube de puntos variables</i>	248
<i>Centro de gravedad de las modalidades de cada variable</i>	249
<i>Centroides</i>	311
<i>Centros de gravedad</i>	424, 425, 429, 461
<i>Chebychev</i>	325, 328, 421, 442
<i>Círculos de correlación</i>	144, 145
<i>City-block</i>	421
<i>Clasificación con dos grupos</i>	458
<i>Clasificación con más de dos grupos</i>	470
<i>Clasificación de las técnicas de análisis multivariante</i>	1, 2
<i>Clasificación de las técnicas de Data Mining</i> .17	
<i>Clasificación global de las técnicas de análisis multivariante de datoS</i>	3
<i>Clasificar individuos futuros</i>	493
<i>Clasificar nuevos casos en que se desconozca el grupo a que probablemente pertenecen</i>	2
<i>Clusters disjuntos</i>	418
<i>Clusters jerárquicos</i>	427, 428, 432
<i>Clusters no jerárquicos</i>	423, 425
<i>Cochrane-Orcutt</i>	66
<i>Coeficiente absoluto de asimetría</i>	37, 38

- Coeficiente de alienación.....* 286
Coeficiente de alienación κ de Guttman 287
Coeficiente de apertura..... 34
Coeficiente de asimetría de Bowley 37
Coeficiente de asimetría de Fisher..... 36
Coeficiente de asimetría de Fisher estandarizado 36
Coeficiente de asimetría de Pearson 37
Coeficiente de contingencia χ^2 359
Coeficiente de contingencia C de K. Pearson..... 361
Coeficiente de correlación por rangos de Spearman. 363
Coeficiente de curtosis 38
Coeficiente de curtosis estandarizado..... 38
Coeficiente de incertidumbre 366
Coeficiente de variación de Pearson 35, 36
Coeficiente de variación intercuartílico..... 34, 36
Coeficiente de variación winsorizado..... 53
Coeficiente R de Pearson..... 366
Coeficiente T de Tschuprow..... 361
Coeficiente V de Cramer 361
Coeficientes de asociación 360, 422
Coeficientes de correlación..... 86, 363
Coeficientes de correlación de Pearson y Spearman. 422
Coeficientes de correlación por rangos de
 Kendall τ_a , τ_b y τ_c 363
Coeficientes de correlación reproducidos..... 208
Coeficientes Lambda de Goodman y Kruskall 362
Comando CONJOINT 607
Componente principal h-ésima.... 127, 128, 142, 233
Componentes de la varianza..... 533, 534
Componentes principales a retener..... 130
Componentes principales categóricas 349
Componentes principales categórico (CATPCA) 279
Componentes principales como caso particular del análisis factorial general 231
Componentes principales y análisis factorial 193
Comprobación de los supuestos del análisis multivariante..... 55
Comprobación de los supuestos subyacentes en los métodos multivariantes 22
Comunalidad 13, 144, 145, 152, 160, 169, 170, 180, 182, 183, 184, 193, 207
Comunalidades y especificidades..... 159
Concepto de análisis cluster..... 417
Concepto de análisis conjunto..... 595
Concepto de análisis discriminante 457
Concepto de escalamiento óptimo..... 275
Concordancia simple 442
Condicionalidad..... 326
Configuración del número de tarjetas de estímulos a generar 612
Configuración inicial de los estímulos..... 285
Conglomerado de k medias..... 437
Conglomerado de pertenencia..... 439, 441, 442
Conglomerados en dos fases..... 452, 453
Conglomerados jerárquicos 441, 450
Contraste de asimetría formal para una variable 37
Contraste de curtosis formal para una variable..... 38
Contraste de esfericidad de Barlett..... 131, 175
Contraste de esfericidad de Barlett 175
Contraste de Kolmogorov-Smirnov Lilliefors de la bondad de ajuste 59
Contraste de normalidad de Shapiro y Wilks 61
Contraste formal estadístico para detectar valores atípicos 51
Contraste para la bondad de ajuste en el método MINRES..... 179
Contraste para la bondad de ajuste en el método ML de máxima verosimilitud..... 177
Contraste sobre las raíces características no retenidas..... 131
Contrastes de la bondad de ajuste 58
Contrastes de significación..... 463, 464, 472
Contrastes después de la obtención de los factores 177
Contrastes en el modelo factorial..... 175
Contrastes formales de heteroscedasticidad..... 64
Contrastes que se aplican previamente a la extracción de los factores 175
Contrastes sobre el número de componentes principales a retener 130
Contrastes y probabilidad de pertenencia 463
Contribución de cada variable a cada función discriminante 486
Contribuciones a la inercia total 271
Convergencia de s-stress 326
Coordenadas del estímulo 292
Coordenadas estímulos 336
Correlación canónica 5
Correlación canónica no lineal 278
Correlación canónica no lineal con SPSS..... 317
Correlación de Pearson..... 45, 442
Correlación Phi de 4 puntos 442
Correlación serial..... 22, 129, 501
Correlaciones bivariadas 446
Correlaciones de variables originales 310
Correlaciones de variables transformadas 310
Correlaciones dicotomizadas..... 45
Covariables 7, 8, 399, 516, 520, 521, 527, 530, 531, 534, 562, 565, 570, 571, 575, 583
Covariantes 514, 515, 540
Crear matriz de distancias..... 325
Criterio de Hotelling 426
Criterio de información bayesiano (BIC)..... 446
Criterio de información de Akaike (AIC) 446
Criterio de la media aritmética..... 130
Criterio de la varianza 221

Criterio de Wilks	426
Cuadrado de la contingencia	359
Cuadrado medio de la contingencia	360
Cuadrado medio del error	500, 507
Cuadrado medio del factor	500
Cuádruple restricción	432
Cuantificaciones de categorías	310

D

D de Anderberg	442
D de Somers	364, 381
D ² de Mahalanobis	55, 420
Data Mining	16, 17
Datos ausentes	46
Datos binarios	267, 325, 327, 422, 442
Definición de distancias	239
Definir el problema de investigación	18
Definir escala	305
Definir escala y ponderación	305
Dendograma	427, 431, 443, 444, 450
Desempatar observaciones empatadas	326
Desviación media respecto de la media aritmética	35
Desviación media respecto de la mediana	35, 53
Desviación respecto de la mediana (MAD)	53
Desviación semintercuartil	34
Desviación típica winsorizada	51, 53
Desviaciones medias	35
Detección bivariante de casos atípicos	54
Detección de valores atípicos	22, 48
Detección gráfica de falta de linealidad	67
Detección multivariante de casos atípicos	55
Detección univariante de casos atípicos	49
Detección y diagnóstico de los datos ausentes	40
Detectar valores atípicos	26, 28, 68, 86
Detector automático de interacción	433
Devaluación media	284
DFITS e Influencia	55
Diagnóstico de los datos ausentes	40
Diagnóstico de los datos ausentes con SPSS	84, 86
Diagrama de control	49
Diagrama de dispersión biespacial	256
Diagrama de dispersión matricial	83
Diagrama de tallo y hojas	25
Diferencia de configuración	325, 327, 442
Diferencia de tamaño	325, 327, 442
Dimensiones comunes o factores	155
Dimensiones en la solución	253, 261, 305
Discretizar	307
Diseño de bloques al azar	510, 511, 512, 513
Diseño de bloques completamente aleatorizados (ANOVA IIF de dos factores fijos sin interacciones)	511
Diseño de medidas repetidas	511, 512
Diseño de tres factores con interacción replicado	543

Diseño de un factor de efectos aleatorios	538
Diseño en cuadrado latino	513
Diseño en parcelas divididas	512
Diseño intrasujeto	511
Diseño Ortogonal	606, 609
Diseño split-splot	512
Diseños completamente aleatorizados	510
Diseños factoriales fraccionados	599, 603
Disparidades	286, 287, 294, 342

Distancia χ^2	220
Distancia cuadrática media de la matriz	293
Distancia cuadrática media por columna	293
Distancia cuadrática media por fila	293
Distancia D ²	55, 420
Distancia de Chebychev	421
Distancia de Mahalanobis	55, 420, 462, 463, 479
Distancia de Manhattan	328, 421
Distancia de Minkowski	421
Distancia de un individuo al centroide de su grupo	421
Distancia desde centro del conglomerado	439
Distancia entre dos individuos	220, 420, 421, 430
Distancia entre las poblaciones	420
Distancia euclídea	220, 221, 222, 223, 232, 241, 293, 295, 301, 325, 326, 327, 328, 419, 425, 428, 430, 438, 439, 442, 462, 463

Distancia euclídea al cuadrado	325, 328, 442
Distancia euclídea de diferencias individuales	326
Distancias entre dos estímulos	292
Distancias y similitudes	419
Distribución de conglomerados	454
Distribución de Weibull	562
Distribución exponencial	561, 562
Distribuciones condicionadas	239, 358
Distribuciones marginales	270, 358, 359, 367, 368, 411, 414
Distribuciones marginales y condicionadas	357
Durbin Watson	66

E

Editor de datos de SPSS	344, 345, 389, 398, 591
Efecto de aprendizaje	512
Efecto de superposición (carry-over effect)	512
Efectos principales	368
Eje factorial h-ésimo	127
Eje factorial q-ésimo	227
Ejes factoriales	241, 244
Eliminación hacia atrás	395
Enfoque de disponibilidad completa	47
Enlace centroide	432
Enlace centroide (centrod method)	429
Enlace completo	431
Enlace completo (complete linkage)	429

<i>Enlace por mínima varianza</i>	430	<i>Estimaciones paramétricas de la función de supervivencia</i>	561
<i>Enlace promedio</i>	430, 432	<i>Estructura factorial de las componentes principales</i> ..	128
<i>Enlace promedio (average linkage)</i>	429	<i>Etiquetar gráficos de las puntuaciones de objeto</i> ... 261, 318	
<i>Enlace simple</i>	430, 431	<i>Etiquetar puntuaciones de los objetos</i>	310
<i>Enlace simple (single linkage)</i>	429	<i>Evaluación de la bondad del ajuste</i>	464, 465
<i>Escalamiento de diferencias individuales</i>	295	<i>Evaluación de los supuestos básicos de la técnica multivariante</i>	18
<i>Escalamiento desdoblado (unfolding)</i>	297	<i>Excluir casos según lista</i>	197
<i>Escalamiento métrico</i>	294, 335	<i>Excluir casos según pareja</i>	197
<i>Escalamiento multidimensional</i> 2, 11, 15, 324, 332, 336, 339, 347		<i>Excluir los valores perdidos</i>	308
<i>Escalamiento multidimensional</i>	281, 291, 324	<i>Excluir objetos con valores perdidos</i>	308, 309
<i>Escalamiento multidimensional (ALSCAL)</i>	324		
<i>Escalamiento multidimensional (EMD)</i>	282		
<i>Escalamiento no métrico</i>	278, 294, 339, 343		
<i>Escalamiento óptimo</i> 261, 271, 303, 305, 317, 318, 349, 353			
<i>Escalamiento para datos de preferencia</i>	297	F	
<i>Escalamiento vectorial</i>	298		
<i>Espacio de las variables R^p</i>	222, 231	<i>Factores comunes</i>	158
<i>Espacio de los individuos Rⁿ</i>	222, 231	<i>Factores no directamente observables</i>	155
<i>Especificidad</i>	159, 160	<i>Factores únicos o factores específicos</i>	158
<i>Esquema general del análisis cluster</i>	436	<i>Fases a seguir en las técnicas de análisis multivariante de datos</i>	18
<i>Esquema general del análisis discriminante</i>	476	<i>Formación de las nubes</i>	239
<i>Esquema general del análisis factorial</i>	194	<i>Fórmula de Lance y Williams para la distancia entre grupos</i>	431
<i>Estadístico de Barlett para contrastación secuencial</i> .472		<i>Fórmula de Lance y Williams para la distancia entre grupos</i>	431
<i>Estadístico de la razón de verosimilitud</i>	583	<i>Frecuencias marginales</i> ... 266, 274, 319, 358, 359, 374	
<i>Estadístico DFITS</i>	55	<i>Función de densidad estimada</i>	560
<i>Estadístico KMO</i>	198, 199, 205, 214, 215, 481	<i>Función de riesgo</i>	560, 561, 562
<i>Estadístico Ra de Rao</i>	465	<i>Función de supervivencia acumulada</i>	560
<i>Estadístico T² de Hotelling</i>	464	<i>Función discriminante de Fisher</i> 458, 462, 468	
<i>Estadístico V de Barlett</i>	465, 472, 473	<i>Función logística</i>	6, 555, 558
<i>Estadísticos basados en distancias</i>	55	<i>Funciones canónicas discriminantes</i>	485, 486
<i>Estadísticos de confianza para puntos de columna</i>	255	<i>Funciones de supervivencia</i>	560
<i>Estadísticos de confianza para puntos de fila</i> ..	255	<i>Funciones lineales discriminantes de Fisher</i>	486
<i>Estadísticos robustos</i>	33	<i>Furthest neighbor</i>	429
<i>Estadísticos robustos centrales</i>	34		
<i>Estadísticos robustos de asimetría y curtosis</i>	36	G	
<i>Estadísticos robustos de dispersión</i>	34, 51		
<i>Estadísticos robustos de la variable</i>	51	<i>Gamma de Goodman y Kruskall</i>	363, 364
<i>Estimación de la media general</i>	376	<i>Generación de un Diseño Ortogonal</i>	606
<i>Estimación de las distancias correspondientes a todos los estímulos</i>	292	<i>Generar diseño</i>	610, 612, 613, 631, 632
<i>Estimación de las utilidades</i>	603	<i>Generar diseño ortogonal</i>	610, 612, 613, 631
<i>Estimación del efecto principal del factor</i>	376	<i>Generate Orthogonal Design</i>	606, 609
<i>Estimación del modelo multivariante</i>	19	<i>Grado de asociación entre variables cualitativas</i>	359
<i>Estimación máximo verosímil de los parámetros del modelo</i>	374	<i>Gráfico biespacial</i>	311
<i>Estimaciones de los parámetros bajo el modelo saturado</i>	393	<i>Gráfico de caja y bigotes</i>	26, 27, 49, 54, 78
<i>Estimaciones no paramétricas de la función de supervivencia</i>	561	<i>Gráfico de caja y bigotes simple</i>	78
		<i>Gráfico de componentes en el espacio rotado</i>	210
		<i>Gráfico de control Individuos y Rango móvil</i>	80
		<i>Gráfico de control tres sigma</i>	50
		<i>Gráfico de dispersión</i>	32
		<i>Gráfico de puntos de fila y columna</i>	271
		<i>Gráfico de sedimentación</i> ... 134, 195, 196, 199, 205, 214, 215	
		<i>Gráfico de simetría</i>	30

Gráfico múltiple de caja y bigotes	28
Gráfico normal de probabilidad.....	56
Gráfico triespacial.....	311
Gráficos de análisis exploratorio con SPSS.....	74
Gráficos de categorías.....	311, 312
Gráficos de control para la detección de casos atípicos.	80
Gráficos de cuantiles.....	77
Gráficos de dispersión	82
Gráficos de grupo.....	326
Gráficos de normalidad	75
Gráficos de probabilidad P-P	75
Gráficos de probabilidad Q-Q	77
Gráficos de transformación	312
Gráficos de variables y objetos	311
Gráficos para los sujetos individuales	326
Grupos más o menos homogéneos en relación al perfil.....	2

H

Herramientas de análisis exploratorio de datos....	49, 54
Heteroscedasticidad	63, 64, 65, 68
Hiperelipsoide de concentración.....	141
Hipótesis de homoscedasticidad.....	463
Hipótesis de normalidad.....	464
Hipótesis en el modelo factorial	158
Histograma de frecuencias.....	23
Histogramas	75
Historial de conglomeración	441
Historial de iteraciones	310, 566
H-medias	425
Homoscedasticidad... 22, 55, 63, 68, 194, 463, 464,	
485, 500	

I

Identificación objetiva	16
Imputación de datos ausentes con SPSS.....	88
Imputación de la información faltante 40, 45, 46, 84,	
108, 110, 112, 116, 117, 118	
Imputación de sustitución por la media.....	47
Imputación de sustitución por la mediana	47
Imputación de sustitución por valor constante....	47
Imputación múltiple.....	48
Imputación por interpolación	47
Imputación por regresión	48
Imputación por sustitución del caso	47
Imputar los valores perdidos.....	308
Independencia condicional.....	373
Independencia parcial: un factor completamente independiente de los demás	373
Independencia total, completa o global.....	373
Independencia y asociación de variables cualitativas.....	359

Independencia y asociación en modelos logarítmico lineales.....	373
---	-----

Índice de asimetría de Kelley	37
-------------------------------------	----

Índice de dispersión respecto a la mediana	35, 36
--	--------

Índice de entropía de Shannon	411
-------------------------------------	-----

INDSCAL (Individual Differences Scaling)	284, 295
--	----------

Inercia de la nube	121, 141, 142, 145, 221, 222, 223, 242
--------------------------	--

Inercia debida a una variable	249
-------------------------------------	-----

Inercia debida a la modalidad	249
-------------------------------------	-----

Inercia explicada por las k primeras componentes principales	128
--	-----

Inercia total	127, 142, 146, 148, 250, 259, 260, 270, 271
---------------------	---

Inercia total de la nube de puntos.....	127
---	-----

Influencia (Leverage)	55
-----------------------------	----

Inspección de los puntos de columna	255
---	-----

Inspección de los puntos de fila	255
--	-----

Interacciones	371, 506
---------------------	----------

Interpretación de las salidas del Análisis Conjunto	624
---	-----

Interpretación de los resultados del escalamiento multidimensional	289, 290
--	----------

Interpretación dimensional	299
----------------------------------	-----

Interpretación geométrica clásica de los componentes principales	141
--	-----

Interpretación geométrica del análisis en componentes principales	138
---	-----

Interpretación geométrica del análisis factorial	179
--	-----

Interpretación por agrupamientos	300
--	-----

Interpretación por regiones	301
-----------------------------------	-----

Interpretación y validación de los resultados	289
---	-----

Interpretar las soluciones MDS	299
--------------------------------------	-----

K

K-medias	424
----------------	-----

KMO de Kaiser, Meyer y Olkin	176
------------------------------------	-----

L

Lance y Williams	325, 327, 431, 432, 442
------------------------	-------------------------

Las medias de filas y columnas	254
--------------------------------------	-----

Leptocúrtica	38
--------------------	----

Linealidad	66
------------------	----

LíneaSp nominal	306
-----------------------	-----

LíneaSp ordinal	306
-----------------------	-----

LíneasSp	306
----------------	-----

Logaritmo del odds ratio	556
--------------------------------	-----

Logaritmos de las razones esperadas de la variable dependiente (odds)	388
---	-----

Logística binaria	564, 570
-------------------------	----------

M

Mapa territorial	479, 491, 492, 493
------------------------	--------------------

Mapas perceptuales	283
--------------------------	-----

Matriz de cargas factoriales.....	144, 146
Matriz de cargas factoriales.....	144
Matriz de componentes	200, 206, 207, 209, 210
Matriz de componentes rotados.....	209
Matriz de coordenadas normalizadas.....	336
Matriz de correlación reproducida.....	161
Matriz de correlaciones anti-imagen.....	205
Matriz de correlaciones reproducidas.....	208
Matriz de inercia.....	124
Matriz factorial ...	160, 164, 165, 172, 173, 174, 181, 183, 184, 185, 188, 199, 214, 217
MDS conjuntamente con otras técnicas	301
MDS de diferencias individuales.....	295
MDS en la investigación.....	302
MDS métrico	295, 297
MDS no-métrico	294
Media truncada	33, 51, 53
Media winsorizada.....	51, 53
Medias de columnas.....	254
Medias de filas	254
Medición de componentes principales.....	189
Medición de los factores mediante el método de Anderson y Rubin	191
Medición de los factores mediante el método de Bartlett.....	191
Medición de los factores mediante estimación por mínimos cuadrados.....	190
Medición de los factores mediante estimación por regresión	190
Medida de chi-cuadrado.....	325, 328
Medida de distancia.....	254
Medida de la información	221
Medida de Phi-cuadrado.....	328, 442
Medida Eta	365
Medida KMO de Kaiser, Meyer y Olkin de adecuación muestral global al modelo factorial	176
Medida MSA de adecuación individual	176
Medida RSQ	286, 287
Medida S-stress utilizada por el algoritmo ALSCAL..	286
Medida Stress de Kruskal	286
Medidas binarias	423
Medidas de asimetría	36, 38
Medidas de asociación direccionales	413
Medidas de asociación globales o simétricas..	413
Medidas de curtosis	36, 38
Medidas de dispersión absolutas no referentes a promedios	34
Medidas de dispersión absolutas referentes a promedios	35
Medidas de dispersión relativas no referentes a promedios	34
Medidas de dispersión relativas referentes a promedios	35
Medidas de la bondad del ajuste más usuales.....	286
Medidas de proximidad o similitud.....	422
Medidas de similaridad para probabilidades condicionales.....	422
Mesocúrtica	38
M-estimador de Andrews	34
M-estimador de Hampel	34
M-estimador de Hubert	34
M-estimador de Tukey	34
M-estimadores	33, 34, 99, 105
Método Alpha para obtener los factores	165
Método asociativo de Williams y Lambert....	432
Método Biquartimax	186
Método Covarimin	187, 188
Método de Anderson y Rubin	191
Método de Bartlett	191
Método de componentes principales iteradas o ejes principales para obtener los factores	169
Método de conglomeración.....	442
Método de estandarización	254
Método de Forgy	424
Método de Fortin	424
Método de imputación de sustitución por la media	47
Método de imputación de sustitución por la mediana	47
Método de imputación de sustitución por valor constante	47
Método de imputación múltiple	47, 48
Método de imputación por interpolación	47
Método de imputación por regresión.....	47, 48
Método de imputación por sustitución del caso	47
Método de la media ponderada (average linkage within groups)	429
Método de la mediana	431
Método de la mediana (median method).....	429
Método de las combinaciones de Wolf	424
Método de las componentes principales para obtener los factores	167
Método de las distancias máximas	429
Método de las distancias mínimas	429
Método de las nubes dinámicas	424
Método de los mínimos cuadrados generalizados GLS.....	174
Método de los mínimos cuadrados no ponderados ULS.....	174
Método de máxima verosimilitud ML	174
Método de máxima verosimilitud para obtener los factores	170
Método de normalización	254, 309
Método de Rotación Promax	188
Método de Taxmap de Carmichael y Sneath.....	424
Método de turstone para obtener los factores	160

Método de Ward.....	430, 442
Método del centroide.....	429
Método del centroide para obtener los factores.....	165
Método del Concepto Completo.....	605, 606, 607
Método del Enlace por densidad.....	430
Método del factor principal para obtener los factores.....	162
Método del límite producto de Kaplan Meier....	561
Método del promedio entre grupos	429
Método Equamax.....	182, 186
Método flexible.....	432
Método ISODATA	427
Método K-means (o K-medias) de mequeen.....	424
Método Oblimax	186
Método Oblimin directo.....	188
Método Ortomax general	186
Método potencial	294
Método Principal por objeto	309
Método Principal por variable	309
Método Quartimax	184
Método Quartimin.....	186, 187, 188
Método Simétrico.....	309
Método Varimax	183
Métodos aglomerativos	429
Métodos centroides.....	424
Métodos de búsqueda de la densidad	424
Métodos de clasificación disociativos	432
Métodos de composición.....	291
Métodos de dependencia	2
Métodos de descomposición	291
Métodos de interdependencia	2
Métodos de reasignación.....	424
Métodos de reducción de dimensiones	424
Métodos del análisis multivariante de la dependencia.....	10
Métodos del análisis multivariante de la interdependencia.....	15
Métodos descriptivos.....	11
Métodos directos.....	170, 424
Métodos disociativos.....	429, 432
Métodos explicativos	4
Métodos MINRES, ULS y GLS para obtener los factores.....	173
Métodos multivariantes de reducción de la dimensión.....	1
Métodos Oblimin	187
Métodos Ortomax	186
Métodos partitivos	423
Minería de datos	16
Minkowski.....	292, 325, 328, 421, 442
MLG univariante y multivariante	531
Modelo ANCOVA.....	7
Modelo ANOVA.....	7
Modelo ANOVA factorial con tres factores	512
Modelo ANOVA II general con efectos aleatorios ...	508
Modelo ANOVA II general con efectos fijos	508
Modelo ANOVA II general con efectos mixtos	509
Modelo ANOVA IIF de dos factores fijos sin interacciones.....	511
Modelo ANOVA IIF general	508
Modelo ANOVA IIM general.....	509, 510
Modelo ANOVA unifactorial de efectos fijos	497
Modelo autorregresivo de orden 1 AR(1)	66
Modelo autorregresivo de orden 2 AR(2)	66
Modelo bifactorial de efectos fijos ANOVA IIF.	503
Modelo bifactorial general con efectos aleatorios ANOVA IIA.....	508
Modelo bifactorial general con efectos mixtos ANOVA IIM	509
Modelo con dos factores y dos covariantes.....	515
Modelo con dos factores y un covariante	514
Modelo con un factor y un covariante	514
Modelo con un solo factor fijo.....	535
Modelo de componentes de la varianza.....	501, 502, 538
Modelo de dos factores y un covariante	514
Modelo de escalamiento de diferencias individuales..	295
Modelo de escalamiento desdoblado (<i>unfolding</i>)	297
Modelo de escalamiento métrico	292
Modelo de escalamiento no métrico	294
Modelo de escalamiento vectorial.....	298
Modelo de MDS métrico.....	294
Modelo de medias móviles de orden 1 MA(1)	66
Modelo de regresión logística....	553, 554, 555, 556, 563
Modelo de regresión Múltiple Lineal General	516
Modelo desdoblado (<i>unfolding</i>).....	297
Modelo en bloques aleatorizados	510
Modelo en cuadrado latino	513
Modelo factorial	158, 175
Modelo general con efectos mixtos	508
Modelo INDSCAL	284, 295, 297
Modelo jerárquico	372
Modelo lineal de probabilidad.....	6, 558
Modelo lineal general (GLM).....	516
Modelo Lineal General en Medidas Repetidas...	548
Modelo Lineal General Multivariante	545
Modelo logarítmico lineal	366
Modelo logístico	553, 554, 555, 556, 557, 558
Modelo Logit	6, 558, 575, 576, 625
Modelo log-logístico	562
Modelo MANCOVA	8, 516
Modelo MANOVA	8, 515
Modelo MDS de diferencias individuales.....	295
Modelo multifactorial de la varianza.....	496, 503
Modelo ponderado	295
Modelo Probit	7, 558, 588
Modelo saturado	372
Modelo unifactorial de efectos aleatorios	501

- Modelo vectorial*.....297, 298, 300
Modelos ANCOVA de la covarianza.....514
Modelos ANOVA II_A general y ANOVA II_M general.....508
Modelos autorregresivos de medias móviles.....66
Modelos con variables cualitativas.....551
Modelos de elección discreta.....6, 551
Modelos de elección discreta con variables ficticias..10
Modelos de escalamiento multidimensional.....291
Modelos de escalamiento óptimo.....275
Modelos de escalamiento para datos de preferencia..297
Modelos euclídeos generalizados.....295
Modelos logarítmico lineales....366, 367, 372, 388, 389
Modelos loglineales.....2, 260, 275, 276
Modelos prácticos de escalamiento multidimensional.....291
Modelos probit y logit.....558, 573
Momento central de orden dos.....430
Monotético.....432
Mostrar el diseño.....613, 614
Mostrar intervalos de tiempo.....577
Multicolinealidad.....65
Múltiples conjuntos.....303, 318, 353

N

- Nearest neighbor*.....429
Nivel de ajuste.....286
Nivel nominal.....277
Nivel numérico (de intervalo).....277
Nivel ordinal.....277
Normalidad.....56, 194, 476
Normalidad en los residuos.....68
Normalización de Kaiser.....146, 152, 153, 183
Número de abandonos.....560
Número de componentes principales a retener...130
Número de fallos en cada intervalo.....560
Número de observaciones expuestas a riesgo560

O

- Objetivo del análisis en componentes principales*....121
Objetivo del análisis factorial155
Objetivo general del análisis factorial223
Objetivos y técnica multivariante conveniente.....18
Objetos y variables311
Objetos, saturaciones y centroides311
Obtención de las componentes principales....123
Obtención de los factores.....248
Odds ratio.....553, 555, 556, 557, 563
Ontrastes de normalidad de asimetría, curtosis y Jarque-Bera.....61
Outliers.....22, 23, 26, 27, 29, 421, 429

P

- Parámetro de escala*.....562
Parámetro de forma.....562
Parámetros poblacionales conocidos.....58
Parámetros poblacionales desconocidos.....59
Pérdida de inercia mínima.....430
Pérdida de memoria.....561
Perfil completo (Full Concept) ..605
Perfil completo (full profile).....603
Perfil de la observación ..239
Perfil de la variable ..239
Perfil de las columnas en Rⁿ ..240
Perfil de las líneas en R^p ..240
Perfiles de fila ..255
Perfiles o porcentajes ..239
Permutaciones de la tabla de correspondencias....255
Pesos o cargas factoriales.....141, 180, 181
Platicúrtica ..38
Polítético ..432
Ponderar casos ..389, 398
Porcentaje de inercia explicada..128, 142, 143, 229, 233
Porcentaje de inercia explicada por la componente principal h-ésima ..128, 233
Porcentaje de inercia explicada por las k primeras componentes principales ..128, 142, 143, 233
Postulado de parsimonia.....183
Prais-Winsten.....66
Predecir el valor de la variable dependiente2
Preferencias en el escalamiento multidimensional. 285, 288
Preparación de tarjetas de estímulos.....613
Primera componente principal.....123, 124, 126
Primera función discriminante.....482, 489
Primera prueba para valorar los datos ausentes. 40, 84, 108
Principio de interpretabilidad.....155
Principio de parsimonia.....155
Principios del análisis cluster ..435
Principios del análisis discriminante ..475
Probabilidad a posteriori de pertenencia a un grupo....466
Probabilidad de pertenencia.....463
Probabilidad de pertenencia a un grupo....463, 466
Probabilidades a posteriori.....466, 467, 469
Probabilidades a priori 466, 467, 468, 469, 479, 486
Probabilidades con información a priori.467, 468, 469
Probabilidades de pertenencia a una población.464, 466
Probabilidades sin información a priori.....467
Procedimiento ALSCAL.....278, 336, 324
Procedimiento Análisis discriminante.....487
Procedimiento Análisis logit lineal.....388
Procedimiento Análisis loglineal general.....398
Procedimiento anova de un factor ..517

<i>Procedimiento Componentes de la varianza</i>	533, 540
<i>Procedimiento común para la determinación de las posiciones óptimas</i>	285
<i>Procedimiento Correlaciones bivariadas</i> ..	86, 89, 446
<i>Procedimiento de MDS no métrico</i>	294
<i>Procedimiento descriptivos de SPSS</i>	92
<i>Procedimiento Escalamiento multidimensional (ALSCAL)</i>	324
<i>Procedimiento Explorar</i>	69, 89, 446
<i>Procedimiento frecuencias de SPSS</i>	90
<i>Procedimiento Generar Diseño Ortogonal</i>	609
<i>Procedimiento KAPLAN-MEIER</i>	580
<i>Procedimiento Loglineal general</i>	388
<i>Procedimiento mlg medidas repetidas</i>	530
<i>Procedimiento MLG multivariante</i>	527
<i>Procedimiento MLG univariante</i>	520
<i>Procedimiento ORTHOPLAN</i>	607
<i>Procedimiento OVERALS</i>	279, 317, 353
<i>Procedimiento PROXSCAL</i>	332
<i>Procedimiento Reemplazar los valores perdidos</i> ..	88
<i>Procedimiento Regresión de Cox</i>	583, 593
<i>Procedimiento resumir</i>	96, 385
<i>Procedimiento Selección del modelo</i>	388, 398
<i>Procedimiento Tablas de contingencia</i> . 251, 381, 412, 446	
<i>Procedimiento tablas de mortalidad</i>	577
<i>Procedimientos informe de estadísticos en filas y columnas de SPSS</i>	94
<i>Proceso completamente aleatorio</i>	40, 84, 112
<i>Propiedades muestrales de las componentes principales</i>	153
<i>Proyección de un punto individuo</i>	248
<i>Proyección de un punto modalidad</i>	248
<i>Proyecto de análisis</i>	18, 19
<i>Prueba ϵ de Ibañez</i>	134
<i>Prueba de Anderson</i>	132
<i>Prueba de Bartlett</i>	198, 481
<i>Prueba de Box</i>	481
<i>Prueba de homogeneidad de varianzas</i>	537
<i>Prueba de las correlaciones dicotomizadas</i> ..40, 108	
<i>Prueba de Lebart y Fenelón</i>	132
<i>Prueba de significación</i>	87
<i>Prueba del bastón roto de Frontier</i>	133
<i>Prueba T para muestras independientes</i> .84, 85, 112	
<i>Pruebas de asociación parcial</i>	390, 393
<i>Pruebas simultáneas</i>	390, 393
<i>Punto de corte discriminante</i>	461, 468, 469
<i>Puntos columna</i>	241, 245, 248
<i>Puntos de columna</i>	256
<i>Puntos de fila</i>	256, 270
<i>Puntos de objetos</i>	311
<i>Puntos dominantes de la solución</i>	260, 271
<i>Puntos línea</i>	245, 248
<i>Puntuación discriminante</i>	458, 466, 467, 490
<i>Puntuaciones de filas</i>	259, 270
<i>Puntuaciones de los objetos</i>	310
<i>Puntuaciones discriminantes</i>	460, 461, 465, 466, 479, 485, 490, 493
<i>Puntuaciones factoriales</i>	195, 197, 208
<i>Puntuaciones o medición de las componentes</i>	129
<i>Puntuaciones o medición de los factores</i>	189
<i>Puntuaciones Z</i>	328, 442
Q	
<i>Q de Yule</i>	442
<i>Quick Cluster Analysis</i>	424
R	
<i>Reconstrucción de la tabla de frecuencias</i>	245
<i>Reconstrucción de la tabla inicial de datos a partir de los ejes factoriales</i>	230
<i>Recorrido intercuartílico</i>	34
<i>Recorrido relativo</i>	34
<i>Recorrido semintercuartílico</i>	34
<i>Reducir la dimensión de una tabla de datos excesivamente grande</i>	2, 11, 12, 474
<i>Reemplazar por la media</i>	197
<i>Regresión de Cox</i>	562, 583, 586, 590, 593
<i>Regresión de Cox con covariable dependiente del tiempo</i>	586
<i>Regresión logística</i>1, 66, 418, 563, 564, 566, 569, 570, 571	
<i>Regresión Logística Binaria</i>	587
<i>Regresión logística multinomial</i>	569
<i>Regresión múltiple</i>	5, 10
<i>Regresión múltiple con variables ficticias</i>	10
<i>Regresión ortogonal y las componentes principales</i> ..	137
<i>Regresión probit</i>	573
<i>Regresión Probit</i>	588
<i>Regresión sobre componentes principales</i>	135
<i>Relación entre el análisis de la regresión y el análisis discriminante</i>	462
<i>Relación entre los análisis en los espacios R^p y R^q</i> ...	228
<i>Relación entre los análisis en R^p y R^q</i>	244
<i>Relacionar todas las variables</i>	2
<i>Relaciones baricéntricas</i>	245
<i>Relaciones entre filas y columnas</i>	245
<i>Replicabilidad</i>	286, 287
<i>Residuos estandarizados</i>	68
<i>Residuos estudiantizados</i>	68
<i>Restricciones para las categorías</i>	253
<i>Retención de variables</i>	134
<i>Rotación de las componentes</i>	145
<i>Rotación de los ejes</i>	146
<i>Rotación de los factores</i>	182

<i>Rotación oblicua</i>	147, 182
<i>Rotación ortogonal</i>	150, 151, 152, 153, 182, 183
<i>Rotación Promax</i>	188
<i>Rotaciones oblicuas</i>	146, 186, 195
<i>Rotaciones ortogonales</i>	146, 152, 186
<i>Rotaciones ortogonales</i>	183

S

<i>Saturaciones en componente</i>	311
<i>Saturaciones y centroides</i>	311
<i>Segmentación jerárquica</i>	9
<i>Segunda componente principal</i>	125, 126
<i>Segunda prueba para valorar los datos ausentes</i>	40, 111
<i>Selección automática del número de conglomerados</i>	445
<i>Selección de variables</i>	464, 466, 477
<i>Selección hacia adelante (forward)</i>	466
<i>Selección hacia atrás (backward)</i>	466
<i>Selección paso a paso (stepwise)</i>	466
<i>Similaridad multivariante</i>	428
<i>Solución habitual para la falta de normalidad</i>	62
<i>Solución, ajuste y preferencias en el escalamiento multidimensional</i>	285
<i>Soluciones más comunes para la multicolinealidad</i>	65
<i>Soluciones para los datos ausentes</i>	46
<i>SPSS y los modelos probit y logit</i>	573
<i>SPSS y correspondencias simples</i>	251
<i>SPSS y el análisis cluster en dos fases</i>	445
<i>SPSS y el análisis cluster jerárquico</i>	440
<i>SPSS y el análisis cluster no jerárquico</i>	437
<i>SPSS y el análisis discriminante</i>	477
<i>SPSS y el análisis factorial</i>	195
<i>SPSS y el escalamiento multidimensional</i>	324
<i>SPSS y el escalamiento óptimo</i>	303
<i>SPSS y el método del concepto completo</i>	606
<i>SPSS y la regresión logística</i>	563, 569
<i>SPSS y la regresión logística multinomial</i>	569
<i>SPSS y las correspondencias múltiples</i>	260
<i>SPSS y los modelos logarítmico lineales</i>	388
<i>Supresión de casos según lista</i>	46, 89, 117
<i>Supresión de datos</i>	46
<i>Supresión de datos según pareja</i>	46, 89, 117
<i>Supresión de los datos ausentes con SPSS</i>	89
<i>Suprimir los casos (filas) o variables (columnas)</i> 46,	
89, 117	
<i>Supuestos estadísticos subyacentes</i>	55
<i>Supuestos subyacentes en los métodos multivariantes</i>	22

T

<i>T² de Hotelling</i>	464, 465
<i>Tabla de Burt</i>	13, 248
<i>Tabla de Burt</i>	248

<i>Tabla de contingencia</i> . 13, 237, 239, 240, 246, 247,	
253, 255, 259, 267, 276, 357, 358, 359, 360, 361,	
362, 367, 372, 374, 375, 377, 378, 379, 381, 389,	
411, 412, 571	

<i>Tabla de contingencia en frecuencias relativas</i> ... 240	
---	--

<i>Tabla de correspondencias</i>	255
--	-----

<i>Tabla de frecuencias observadas, esperadas y</i>	
<i>residuales</i>	392, 397

<i>Tabla disyuntiva completa</i>	247
--	-----

<i>Tablas de contingencia bidimensionales</i>	374
---	-----

<i>Tablas de frecuencias marginales</i>	265, 273
---	----------

<i>Tablas de mortalidad</i>	560, 577, 589
-----------------------------------	---------------

<i>Tarjetas de estímulos</i>	606, 612, 621
------------------------------------	---------------

<i>Técnica multivariante conveniente</i>	18
--	----

<i>Técnicas auxiliares</i>	16, 17, 299
----------------------------------	-------------

<i>Técnicas composicionales y descomposicionales</i> ...	598
--	-----

<i>Técnicas de modelado originado por la teoría</i> ..	16
--	----

<i>Técnicas de modelado originado por los datos</i>	16, 17
--	--------

<i>Técnicas del análisis de la interdependencia</i>	11
---	----

<i>Técnicas del análisis de la dependencia</i>	4
--	---

<i>Técnicas emergentes de análisis multivariante de datos</i> ..	16
--	----

<i>Técnicas multivariantes analíticas o inferenciales</i> ..	2
--	---

<i>Técnicas multivariantes descriptivas</i>	2
---	---

<i>Teoría de la integración de la información</i>	601
---	-----

<i>Test conjunto de aleatoriedad de Little</i>	40
--	----

<i>Test de BARLETT</i>	500
------------------------------	-----

<i>Test de BONFERRONI</i>	501
---------------------------------	-----

<i>Test de Dixon</i>	51, 53
----------------------------	--------

<i>Test de esfericidad de Barlett</i>	205
---	-----

<i>Test de Grubbs</i>	51
-----------------------------	----

<i>Test de HARTLEY</i>	500
------------------------------	-----

<i>Test de KOLMOGOROV-SMIRNOV</i>	500
---	-----

<i>Test de la chi-cuadrado</i>	13, 58
--------------------------------------	--------

<i>Test de la diferencia mínima significativa</i>	501
---	-----

<i>Test de recorrido múltiple de DUNCAN</i>	501
---	-----

<i>Test de SCHEFFE</i>	501
------------------------------	-----

<i>Test de TUKEY de comparaciones múltiples</i>	501
---	-----

<i>Test HSD de TUKEY</i>	501
--------------------------------	-----

<i>Test Q de COCHRAN</i>	500
--------------------------------	-----

<i>Test SNK de STUDENT_NEWMAN_KEULS</i>	501
---	-----

<i>Test W de SHAPIRO y WILK</i>	500
---------------------------------------	-----

<i>Tipos de gráficos</i>	74
--------------------------------	----

<i>Transformaciones algebraicas</i>	39
---	----

<i>Transformaciones de las variables</i>	39
--	----

<i>Transformaciones lineales</i>	39
--	----

<i>Transformaciones lógicas</i>	39
---------------------------------------	----

<i>Transformaciones no lineales no monotónicas</i> ..	39
---	----

<i>Tratar distancias menores que n como perdidas</i> ..	326
---	-----

U

<i>Unicidad de las soluciones</i>	161
---	-----

V

<i>Validación de los resultados</i>	290
<i>Validación del modelo multivariante</i>	19
<i>Valor mínimo de s-stress</i>	326
<i>Valoración del ajuste del modelo</i>	19
<i>Valores atípicos</i>	48
<i>Valores ausentes con SPSS</i>	84
<i>Valores perdidos</i>	197, 308, 438
<i>Variables clasificadoras</i>	437, 441, 458, 462, 464, 465, 466, 472
<i>Variables concomitantes</i>	514
<i>Variables de etiquetado</i>	305
<i>Variables para gráficos biespaciales y triespaciales</i> ..	311
<i>Variables suplementarias</i>	305
<i>Varianza explicada</i>	310
<i>Varianza total</i>	13, 141, 142, 162, 164, 184, 193, 194, 199, 205, 221, 222, 229, 350, 425, 430

<i>Varianzas de las componentes</i>	127
<i>Vecino más lejano</i>	442
<i>Vecino más próximo</i>	442
<i>Vecinos más cercanos</i>	429
<i>Vecinos más lejanos</i>	429
<i>Vinculación inter-grupos</i>	442
<i>Vinculación intra-grupos</i>	442
<i>Visualizar diseño experimental</i>	613

W

<i>Ward</i>	430, 432, 436
-------------------	---------------

Y

<i>Y de Yule</i>	442
------------------------	-----



Técnicas de Análisis Multivariante de Datos Aplicaciones con SPSS® Pérez

El objetivo de este libro es proporcionar una visión clara conceptual de las técnicas estadísticas multivariantes de análisis de datos, describir pormenorizadamente sus principales métodos e ilustrar con ejemplos prácticos su aplicación en los distintos campos de la investigación.

En un primer bloque de contenido se presentan las técnicas de análisis exploratorio de datos para el estudio de grandes conjuntos de datos, tratamiento de los valores atípicos e imputación de datos ausentes.

A continuación se abordan las técnicas más importantes de reducción de la dimensión como son los métodos factoriales (análisis en componentes principales, análisis factorial, análisis de correspondencias, etc.).

El siguiente bloque de contenido se encarga de las técnicas de escalamiento óptimo y multidimensional con sus aplicaciones, seguido de la presentación de las técnicas de clasificación y segmentación mediante análisis cluster y discriminante.

Más adelante se tratan técnicas específicas como los modelos logarítmico lineales, la regresión logística, los modelos Logit y Probit, los modelos del análisis de la varianza y la covarianza y la regresión de Cox y el análisis de la supervivencia.

Por último, se aborda la técnica multivariante del análisis conjunto de datos y sus aplicaciones.



Evidentemente, el objetivo de esta obra no se consigue sin el uso de medios informáticos. Por ello, adicionalmente, en este libro se utiliza uno de los programas de computador más usuales en este campo, concretamente SPSS, y se resuelven los problemas con este software. Asimismo, se acompaña un CD-ROM que contiene los archivos relativos a los ejemplos prácticos del libro.

PEARSON
Educación

www.pearsoneducacion.com

ISBN 978-84-832-2901-9



9 788483 229019