www.flacsoandes.edu.ec

INTIYAN
EDICIONES CIESPAL
521

Métodos de investigación social

Paulina Salinas Meruane
Manuel Cárdenas Castro

Quito - Ecuador
2009

Métodos de investigación social

Primera Edición

© 2008, Ediciones Universidad Católica del Norte
AV. Angamos 0610,Antofagasta, Chile
Telefax:(56)(55)355824 / 355826
E-mail: www.periodismo.ucn.cl
ISBN: 978-956-287-266-9

Segunda Edición

© Paulina Salas Meruane
Manuel Cárdenas Castro
1.000 ejemplares - Marzo 2008

ISBN: 978-9978-55-070-0
Código de barras 978-9978-55-070-0
Registro derecho autoral Nº 030584

Portada y Diagramación
Diego Acevedo

Impresión
Editorial "Quipus", CIESPAL
Quito-Ecuador

Los textos que se publican son de exclusiva responsabilidad de su autor.

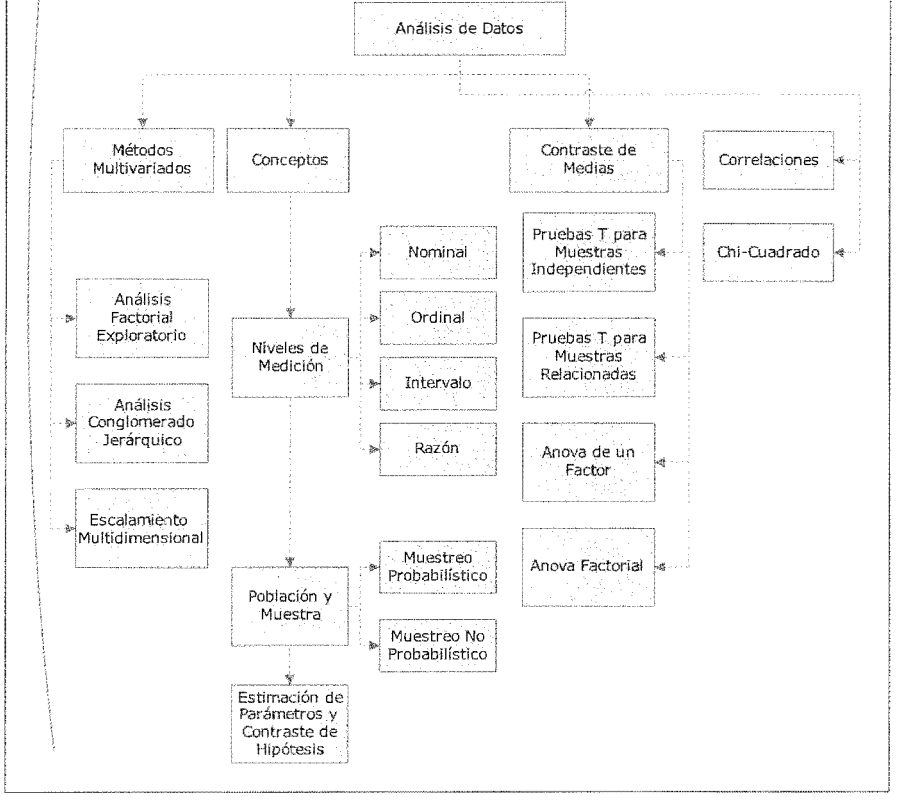
ÍNDICE

Primera Parte
Diseños de Investigación Cuantitativa

Segunda Parte
Diseños de Investigación Cualitativa

CAPITULO 8

Análisis de Datos



Capítulo 8

Procedimientos y Técnicas de Análisis de Información en SPSS 14.0

Manuel Cárdenas Castro

En este capítulo presentamos los principales procedimientos de análisis de datos y su respectiva aplicación en el programa estadístico SPSS. Se realiza una breve descripción de los conceptos fundamentales para posteriormente abordar algunos de los principales procedimientos al uso en ciencias sociales, de modo de distinguir cuándo resulta pertinente realizar el procedimiento (para qué tipo de datos), cómo realizar las operaciones necesarias para su ejecución y el modo de interpretar los análisis. Finalmente, se revisarán algunas de las principales técnicas de análisis multivariados.

Palabras clave: *Análisis de datos, estadística descriptiva e inferencial, métodos multivariados.*

8.1. Introducción

El análisis estadístico de datos engloba un conjunto de procedimientos diseñados para seleccionar datos, describirlos, y extraer conclusiones de ellos (Pardo y San Martín, 2001). Es decir, se Intenta obtener conclusiones relevantes a partir de datos empíricos, y todo ello mediante el uso de modelos matemáticos. Se trataría de un saber organizado que pretende recoger, ordenar

y analizar los datos de una muestra (representativa de una determinada población), para realizar inferencias acerca de dicha población.

En términos generales se suele distinguir entre la estadística descriptiva y la inferencial. La primera de ellas consiste en una serie de procedimientos que tienen como objetivo organizar y resumir la información contenida en un conjunto de datos. La estadística inferencial, por su parte, consiste de una serie de procedimientos que permiten generalizar las propiedades de un conjunto de datos a un conjunto mayor de datos. Es decir, se pretende hacer inferencias válidas sobre una población a partir de los datos obtenidos en una muestra representativa de dicha población (la representatividad de la muestra queda asegurada mediante las técnicas de muestreo, las que pueden revisarse en el capítulo de este manual dedicado al muestreo).

El análisis de datos tiene como objetivo central encontrar relaciones generales (leyes) que expliquen el comportamiento de uno o varios eventos (Pardo y San Martín, 2001). Estas leyes solo pueden ser descubiertas y verificadas observando el mundo real por medio de un método replicable que permita obtener resultados consistentes en condiciones similares.

En el capítulo anterior afirmábamos que medir consistía en asignar números, ahora agregamos que cada uno de esos números corresponde a un dato. Es por ello que, para analizar datos debemos asignar números a las características de las personas u objetos, de manera de conectar dos sistemas de relaciones: uno empírico (el de las propiedades que se desea medir) y otro formal (el de los números que se asignan). Cuando el sistema formal refleja al empírico, entonces dicha correspondencia implica una medición (Pardo y Ruiz, 2002). Hay mediciones mejores que otras, en sentido que en algunos casos se podrá establecer un mayor número de relaciones.

Dependiendo de la riqueza de esas relaciones existirán diferentes niveles o escalas de medida. Tradicionalmente se han distinguido cuatro tipos de escala o niveles de medida: nominal, ordinal, de intervalo y de razón (las que aquí solo reseñaremos brevemente, ya que han sido analizadas en otro capítulo).

- *Las medidas nominales* Son aquellas en las que se clasifica a los sujetos u objetos sobre la base de categorías, de modo que todos los sujetos ubicados en una categoría son similares o equivalentes respecto de la propiedad medida. Posteriormente, se asignan números a dichas categorías. Las categorías utilizadas deben cumplir con los criterios de exhaustividad (a todos los sujetos corresponde una categoría) y exclusividad (cada sujeto solo pertenece a una categoría). Los números asignados funcionan solo como rótulos, por lo que la única relación que se puede establecer es la de igualdad o desigualdad (un ejemplo típico de este nivel es la variable sexo).

- *Las medidas ordinales* Este nivel de medición nos permite ordenar los elementos según la cantidad que poseen de la variable. Permite establecer relaciones de diferencia de cantidad (mayor que o menor que) entre los sujetos y ordenarlos según dicha diferencia. La limitante de este nivel es que aún no se puede afirmar nada respecto de la magnitud de la diferencia entre sujetos (un ejemplo de este nivel de medida es la estatura de los sujetos, cuando desconocemos su medida en centímetros y simplemente podemos ordenarlos de más bajo a más alto).

- *Las medidas de intervalo* Este nivel aporta respecto del anterior la posibilidad de determinar la magnitud de la diferencia en la variable de interés. Eso sí, en esta medida carecemos del cero absoluto (ausencia de la variable), por lo que no podemos afirmar que un evento contenga el doble de la variable simplemente porque los números así lo

indiquen (un ejemplo clásico es el de la temperatura, donde la magnitud de 0º no implica ausencia de temperatura).

- *Las medidas de razón* En este nivel se subsanan los problemas anteriores, ya que aquí el cero ya no es arbitrario sino un punto fijo que implica ausencia de la variable medida. Es justamente por esto que se puede afirmar si un objeto posee el doble o el triple de cantidad de variable que el otro (ejemplos recurrentes para ilustrar este nivel de medición son el tiempo o peso, ya que aquí sí es posible constatar la ausencia de la variable).

La importancia de una adecuada distinción y manejo de los diferentes niveles de medición radica en que la utilización de técnicas de análisis de datos se encuentra siempre mediatizada por el tipo de variable de que se dispone. En todo caso, debemos saber que existe una multitud de variables en las que resulta muy difícil determinar a qué nivel de medida corresponden, ya que se trata de mediciones "subjetivas" (ej. escala de dolor) en las que no se pueden considerar como equivalentes las asignaciones realizadas por diferentes sujetos. En todo caso, debemos recordar que una vez asignados los números todas las técnicas de análisis de datos que revisaremos en este capítulo son posibles de ejecutarse.

Hemos afirmado al comenzar esta revisión que el análisis de datos pretende, entre otras cosas, realizar inferencias válidas para una población a partir de una muestra de esta. Llamamos población o universo al conjunto total de elementos que poseen una característica específica en común y que se desea estudiar. Existen, de acuerdo al número de elementos que conforman el universo, poblaciones finitas (número delimitado de elementos) e infinitas (número ilimitado de elementos). Las poblaciones con las que tiene sentido trabajar suelen ser finitas, pero la mayor parte de las veces son tan grandes que para efectos de análisis pueden ser consideradas como infinitas (Pardo y Ruiz, 2002).

Por su parte, la muestra refiere a un subconjunto representativo de elementos, todos ellos pertenecientes a una determinada población. Son de un tamaño limitado que las hace propicias para trabajar sobre ella y extraer conclusiones referidas a todos los elementos de la población (Pardo y Ruiz, 2002). Las características de dicha muestra serán descritas por medio de estadísticos (media, mediana, varianza, etc.) que representarán los valores concretos poblacionales, los cuales nos son desconocidos (parámetros). La representatividad de la muestra queda asegurada mediante un adecuado procedimiento de selección de los participantes de una muestra (ver capítulo sobre muestreo) que nos garantice que cualquier elemento de la población pudo haber estado representado en dicha muestra. El muestreo puede ser de tipo probabilístico (la probabilidad asociada a la selección de un elemento para formar parte de una muestra puede ser calculada) o no probabilístico (se desconoce o no se toma en cuenta la probabilidad asociada a cada una de las muestras posibles).

Los tipos de muestreo más utilizados son tres: aleatorio sistemático (se asignan números a cada uno de los elementos de una lista, se define el tamaño de la muestra y se calcula la constante dividiendo el total de la población por el de la muestra, luego se selecciona al azar un número incluido en la constante y luego se le suma la constante hasta completar la muestra); *aleatorio estratificado* (cuando población está formada por diferentes subpoblaciones se seleccionan primero los estratos y luego se procede como en el modelo anterior pero resguardando el tamaño de los estratos –afijación proporcional) y *aleatorio por conglomerados* (para casos en que la unidad muestral no son elementos individuales sino grupos. Tiene la ventaja de que no se necesita el listado de todos los elementos de un grupo).

Hemos venido afirmado que el objetivo del análisis de datos es extraer conclusiones generales y realizar predicciones sobre el comportamiento de ciertas variables en la población. Las inferencias pueden realizarse mediante dos estrategias: *estimación*

de parámetros y *contraste de hipótesis*. Ambas estrategias permiten llegar a las mismas conclusiones, pero la información que entregan es algo diferente: el contraste de hipótesis pone el énfasis en intentar detectar la presencia de un efecto significativo y la estimación de parámetros pone énfasis en cuantificar el tamaño del efecto detectado (Pardo y San Martín, 2001; Pardo y Ruiz, 2002).

El contraste de hipótesis, también llamado prueba de significación, es un método que permite tomar decisiones y afirmar con alto grado de certeza si una afirmación acerca de una población puede ser mantenida o si por el contrario debe ser rechazada. Para ello se formulan hipótesis científicas, las que posteriormente son transformadas en hipótesis estadísticas (afirmaciones sobre una o más distribuciones de probabilidad o sobre el valor de uno o más parámetros de esas distribuciones) y se explicita una regla de decisión, la cual se establece en términos de probabilidad (que servirá como criterio para decidir si la hipótesis nula planteada debe o no ser rechazada). De este modo, el contraste de hipótesis es un proceso de decisión en el que una hipótesis formulada en términos estadísticos es puesta en relación con los datos empíricos para determinar si es o no compatible con ellos.

Por otra parte, la estimación de parámetros consiste en utilizar la información muestral para inferir alguna propiedad de la población. Es decir, se utiliza un estadístico (llamado estimador) para inferir el valor de algún parámetro poblacional.

Una vez revisados estos conceptos básicos corresponde pasar a revisar los procedimientos de análisis de datos, para lo cual avanzaremos de forma sucesiva en los procedimientos de comparación de medias (pruebas t para muestras independientes y relacionadas, análisis de varianza de un factor y análisis de la varianza factorial), correlaciones (coeficientes de correlación de Pearson y Spearman) y análisis no paramétrico para variables categóricas (chi-cuadrado). También nos referiremos a algunos

de los métodos multivariantes más utilizados (análisis factorial, análisis de conglomerados y escalamiento multidimensional). A continuación presentamos un tabla con el resumen de los procedimientos (Tabla 1), la que indica al tipo y el número de variables que utiliza (niveles de medición), así como sus características y usos principales, como también las hipótesis que contrasta (cuando resulta pertinente indicarlo)

Procedimiento	Tipo de variables	Características
Pruebas T para muestras independientes	La variable dependiente es una variable cuantitativa (Ej. puntuación en una escala) y la variable independiente es una variable categórica (nominal u ordinal) con dos niveles (Ej. Sexo dicotomizada en hombres y mujeres).	Contrasta la hipótesis nula de que las medias de la dos poblaciones son iguales (u1 = u2). Es decir, no se espera encontrar diferencias significativas entre las dos puntuaciones.
Pruebas T para muestras relacionadas	Se utiliza como variables dos puntuaciones cuantitativas (ej. dos medidas en una prueba) para un mismo grupo de sujetos.	Contrasta la hipótesis nula de igualdad de medias. Es decir, no se espera encontrar diferencias significativas entre las dos puntuaciones.
Análisis de varianza (ANOVA de un factor)	Se utiliza como variable independiente una variable categórica de más de dos niveles (Ej. Nivel socioeconómico = alto, medio y bajo). La variable dependiente es una variable cuantitativa (Ej. Salario).	Contrasta la hipótesis nula de igualdad de medias, por lo que se pone a prueba la hipótesis que afirma que no existen diferencias significativas entre las medias de los grupos formados por la variable categórica.
Análisis de varianza factorial (ANOVA factorial)	Aquí se utilizan dos variables independientes categóricas (Ej. Sexo y Nivel socioeconómico) y una dependiente de tipo cuantitativo (Ej. Salario).	La hipótesis es de igualdad de medias, pero nos permite apreciar la forma en que interactúan las variables (Ej. ver si ser mujer y de un determinado NSE es importante para determinar los niveles de ingresos).
Chi cuadrado	Establece relación y grado de relación entre variables de tipo categórico (Ej. ser mujer y pertenecer a un determinado NSE).	Contrasta hipótesis nula de que los criterios de clasificación son independientes entre sí.
Correlaciones (regresión lineal)	Analiza el grado de variación conjunta entre dos variables cuantitativas (Pearson) o entre variables ordinales (Spearman).	La hipótesis nula es de independencia lineal. Es decir, las variables no estarían relacionadas de forma lineal.

Análisis factorial (exploratorio y confirmatorio)	Se trata de un método de agrupación de datos que permite reducir la complejidad de estos y apreciar factores comunes. Supone todas las variables como independientes, ya que no existe a priorí una dependencia conceptual entre ellas.	Permite explorar datos (AFE) y reducir su dimensionalidad, agrupando variables en factores comunes (realiza sólo una evaluación global). El AFC (confirmatorio) permite contrastar hipótesis sobre un modelo factorial preciso y bien especificado.
Análisis de conglomerados	Procedimiento de agrupación de datos en grupos de variables de acuerdo a su parecido. Se puede especificar el número de conglomerados (k medias) o se puede liberar su obtención (jerárquicos).	Permite agrupan casos o variables en función de su parecido o similitud en un número óptimo de grupos, teniendo en consideración sólo criterios internos (distancias).
Escalamiento multidimensional	Es una técnica multivariada de independencia que busca encontrar información sobre la estructura subyacente o latente de un conjunto de datos, de forma de simplificar su interpretación.	Es útil para determinar las imágenes subjetivas asociadas a un conjunto de objetos por parte de una muestra de sujetos, así como para determinar las dimensiones sobre las cuales basan esos juicios.
Análisis de correspondencias (simples y múltiples)	Se utiliza para determinar la relación o conexión recíproca existente entre un número amplio de variables.	Profundiza en las relaciones de dependencia que se establecen entre variables cualitativas observadas en una misma población (entre dos variables si se trata de AC simple y entre más de dos si se trata de un AC múltiple). Busca la estructura oculta a un conjunto amplio de datos o variables.

8.2. Pruebas T para muestras independientes

Permite contrastar hipótesis sobre diferencia de medias entre dos muestras independientes. Es decir, nos permite saber si existen diferencias significativas en las medias de dos muestras que nos autoricen para afirmar que pertenecen a diferentes poblaciones. La prueba T tiene dos versiones diferentes, dependiendo si la varianza de la población pueden ser consideradas iguales o no. Para dirimir este asunto contamos con la prueba de Levene (estadístico de Levene) que permite contrastar la hipótesis nula sobre igualdad de varianzas.

Para realizar un procedimiento de pruebas T para muestras independientes debemos contar con dos variables: una cuantitativa (dependiente) y otra categórica (independiente) que tenga solo dos niveles. Posteriormente pulsamos el menú Analizar > Comparar medias > Pruebas T para muestras independientes. Una vez realizado el procedimiento se desplegará el cuadro de diálogo correspondiente (Figura 1).

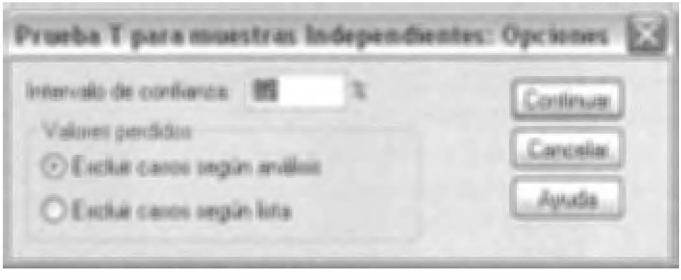
Figura 1. *Prueba T para muestras independientes*

Una vez abierto dicho cuadro de diálogo, trasladamos desde el listado de variables aquellas que nos servirán como variable de agrupación y aquella que nos interesa contrastar. A la variable de agrupación le llamamos variable independiente (y debe ser una variable categórica de dos niveles. Ej. Sexo) y a la variable a contrastar le llamamos variable dependiente (la cual debe ser de tipo cuantitativo. Ej. Media de puntuaciones en una escala).

Si observamos el cuadro de diálogo veremos que nos ofrece activo un botón denominado Opciones. Si pulsamos dicha opción se desplegará el subcuadro de diálogo correspondiente (Figura 2) y en el cual podremos definir el nivel de confianza con el cual deseamos trabajar. La opción por defecto ofrecida por SPSS es del 95%, que significa que nuestra medición contendrá un error del 5%.

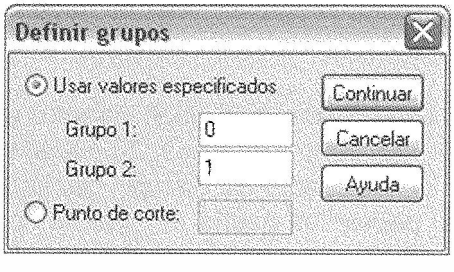
Por otra parte, en el subcuadro Opciones nos permitirá tomar decisiones respecto de qué hacer con los sujetos que no tienen asignado un valor en las variables que trabajamos. En tal caso podemos optar por excluirlos según análisis (se elimina el caso solo para el análisis en curso) o según lista (se excluye de todos los análisis).

Figura 2. *Prueba T: opciones*



Una vez fijado el intervalo de confianza y trasladadas las variables dependiente e independiente a sus respectivas ventanas, debemos especificar los valores que representan a cada dimensión de la variable categórica que utilizamos como criterio de comparación (Figura 3). Aquí debemos respetar los niveles que fijamos al momento de crear la variable en la vista de variables de SPSS (Ej. hombres = 0 y mujeres = 1).

Figura 3. *Prueba T: Definir grupos*



Lo anterior nos permite realizar comparaciones en aquellas variables en que hay más de dos niveles, aunque esto siempre

deba ser hecho por pares de variables. Basta simplemente indicar entre qué grupos se quiere realizar la comparación. Por otra parte, podemos observar cómo el programa también nos ofrece la opción de especificar un punto de corte que divida la muestra en dos grupos (Ej. la mediana, la media, o la puntuación teórica media de la escala, etc.).

Una vez que se ha realizado el procedimiento anteriormente descrito, volvemos al cuadro de diálogo principal y pulsamos la opción aceptar. Inmediatamente se desplegará ante nosotros el visor de resultado, el cual nos mostrará los procedimientos y cálculos que se han realizado a partir de las medias de los dos grupos formados por la variable sexo. Lo primero en aparecer (Figura 4) es el cuadro con el resumen de los estadísticos de grupo, el que nos informa del tamaño de cada grupo (columna N) y sus respectivas medias de respuesta para la escala, acompañadas de su correspondiente desviación típica (que como ya sabemos nos indica el grado de dispersión de los puntajes en torno de la media).

Figura 4. *Prueba T para muestras independientes: Estadísticos descriptivos*

Estadísticos de grupo

	Sexo del sujeto	N	Media	Desviación típ.	Error típ. de la media
PUNTAJE	Mujer	92	19,88	7,595	,792
	Hombre	50	23,72	10,414	1,473

Podemos apreciar en la tabla que resume el procedimiento realizado (figura 5) que en las dos primeras columnas se nos entrega el valor del estadístico de Levene y su respectivo nivel de significación. Este contraste es el que nos permitirá decidir si podemos o no asumir varianzas iguales para las poblaciones. Si la probabilidad asociada es menor a 0.05 se rechazará la hipótesis nula de que las varianzas poblacionales son iguales. En nuestro

ejemplo rechazamos dicha hipótesis debido a que el estadístico tiene asociada una significación de 0.03, es decir, menor a 0.05. Esto nos obligará a leer todos los datos siguientes en la segunda fila, frente al texto "No se han asumido varianzas iguales".

Figura 5. *Prueba T para muestras independientes: Resumen del procedimiento*

Estadísticos de grupo

	Sexo del sujeto	N	Media	Desviación tip.	Error tip. de la media
PUNTAJE	Mujer	92	19,88	7,595	,792
	Hombre	50	23,72	10,414	1,473

En las columnas siguientes encontramos el valor tanto para el estadístico t (-2,296), sus grados de libertad asociados, el nivel crítico asociado o significación (0,24), la diferencia entre las medias (-3,84), el error típico de esa diferencia (1,672) y los límites inferior (-7,169) y superior (-0,511) para el intervalo de confianza. Como al valor del estadístico t está asociada una significación de 0.024 podemos rechazar la hipótesis nula de igualdad de medias. Así, podemos afirmar que los hombres tienen una media significativamente superior a las mujeres en la escala que mide prejuicio. Es decir, se rechaza la hipótesis nula de que las puntuaciones de hombres y mujeres en la escala de prejuicio son iguales.

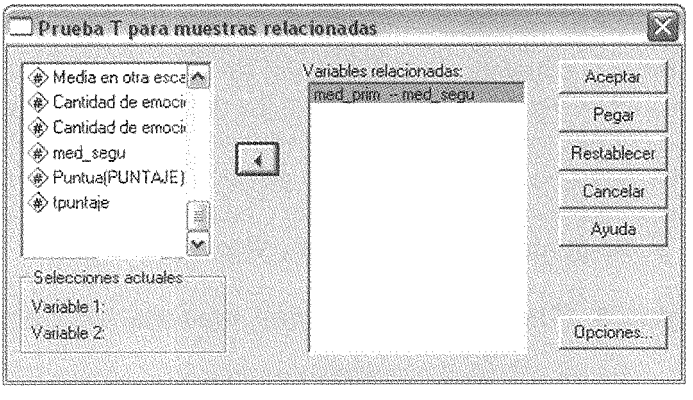
8.3. Pruebas t para muestras relacionadas

Esta prueba es una variación del procedimiento anterior y nos permite contrastar la hipótesis de igualdad de medias entre dos muestras de puntuaciones relacionadas. Lo que hacemos en este caso es trabajar sobre un par de puntuaciones tomadas sobre el mismo grupo

Para realizar un procedimiento de pruebas T para muestras relacionadas debemos contar con dos variables cuantitativas

(dependientes) que representan dos puntuaciones de los sujetos (Ej. los puntajes para un pre-test y un post-test) para un mismo grupo de sujetos. En nuestro caso realizaremos las comparaciones con la muestra completa, pero si quisiéramos realizar la comparación de medias solo para un determinado grupo (Ej. las mujeres), deberíamos seleccionarlo en el menú "Datos" o segmentar el archivo, utilizando como variable de segmentación el sexo de los sujetos. Posteriormente pulsamos el menú Analizar > Comparar medias > Pruebas T para muestras relacionadas. Una vez realizado el procedimiento se desplegará el cuadro de diálogo correspondiente (Figura 6). En dicho cuadro de diálogo debemos traspasar las variables por pares a la ventana "Variables relacionadas" desde la lista de variables (en nuestro ejemplo hemos traspasado las variables "med_prim" y "med_sugu", las que corresponden a dos puntuaciones distintas sobre una misma escala).

Figura 6. *Prueba T para muestras relacionadas*



Las opciones que aparecen en el cuadro de diálogo son similares a las del procedimiento pruebas t para muestras independientes (Figura 2), es decir, podemos controlar el nivel de confianza y decidir qué hacer con los valores perdidos. Una vez realizadas estas operaciones, pulsamos en aceptar y ya podemos encontrar nuestros resultados en el visor de resultados del programa.

Al igual que en el procedimiento anterior, lo primero que nos entrega SPSS es el cuadro que contiene los estadísticos descriptivos para cada aplicación (media y desviación típica), así como la información referida al número de sujetos que se han incorporado en el análisis (Figura 7).

Figura 7. *Prueba T para muestras relacionadas: Estadísticos descriptivos*

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	MED_PRIM	2,3694	142	,92940	,07799
	MED_SEGU	2,1232	142	,88491	,07426

Desde ya podemos apreciar que la media de respuestas en la escala de prejuicio para la primera aplicación ha sido superior que para la segunda. También sabemos que la dispersión de los puntajes respecto de la media sigue la misma lógica. En la Figura 8 podemos apreciar una novedad respecto del procedimiento anterior y consiste en la inclusión en los resultados del cálculo del coeficiente de correlación para los dos grupos de puntuaciones y su correspondiente nivel de significación estadística. Para el caso sabemos que la correlación es positiva y alta, es decir, las dos puntuaciones están relacionadas y cuando se puntúa alto en una se suele puntuar alto en la otra y viceversa.

Figura 8. *Prueba T para muestras relacionadas: Correlaciones*

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	MED_PRIM y MED_SEGU	142	,818	,000

Finalmente, la última información que arroja el programa es la tabla con el resumen del procedimiento (Figura 9). La primera

columna nos entrega la diferencia entre las medias y sus descriptivos asociados. Las dos columnas siguientes nos informan que podemos estimar con un 95% de confianza que la verdadera diferencia entre las medias de las dos aplicaciones de nuestra escala se encuentra entre 0,155 y 0,337. Finalmente, en las tres columnas finales, se nos entrega el valor del estadístico t de Student (5,337), sus respectivos grados de libertad y su nivel de significación. En este caso, el nivel crítico bilateral es menos a 0.05, por lo que podemos rechazar la hipótesis de igualdad de medias y concluir que el promedio de la segunda aplicación es significativamente menor al de la primera, lo que implica que nuestra muestra responde de forma más prejuiciosa en la primera aplicación (Ej. en caso que se hubiese realizado un diseño experimental podríamos haber constatado la utilidad del tratamiento experimental para reducir el prejuicio).

Figura 9. *Prueba T para muestras relacionadas: Resumen del procedimiento*

Prueba de muestras relacionadas

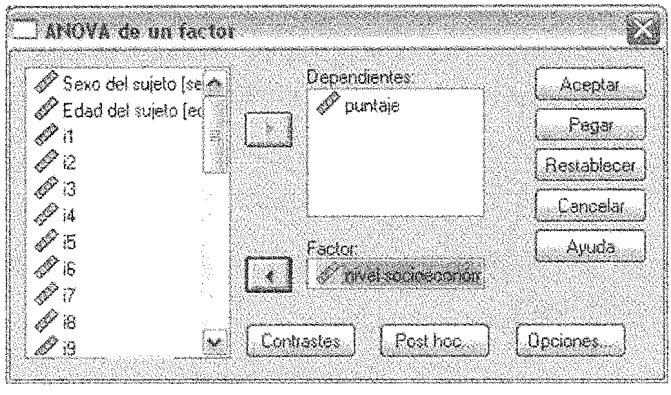
	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia: Inferior	Superior			
Par 1 MED_PRIM - MED_SEGU	,2461	,54364	,04612	,1550	,3373	5,337	141	,000

8.3. Análisis de Varianza (ANOVA) de un Factor

El procedimiento de análisis de varianza de un factor nos permite comparar a varios grupos en una medida cuantitativa. Es un análisis similar al de pruebas t para muestras independientes, pero útil cuando los niveles de la variable independiente (factor) son superiores a dos. La hipótesis que se pone a prueba en este análisis es de igualdad de medias poblacionales, y el estadístico de contraste se representa con la letra F (el valor asociado a este estadístico será más elevado mientras mayor sean las diferencias de medias entre los grupos comparados).

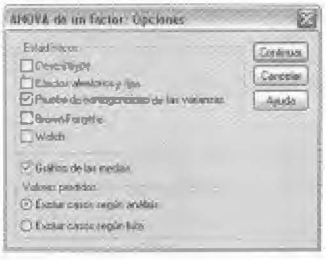
Para realizar este procedimiento debemos pulsar el menú Analizar > Comparar medias > ANOVA de un factor. Una vez realizado el procedimiento se desplegará el cuadro de diálogo correspondiente (Figura 10). Una vez en él, debemos trasladar la variable cuantitativa (VD) a la ventana "Dependientes" y la variable categórica (VI) de más de dos niveles a la ventana "Factor".

Figura 10. *ANOVA de un factor*



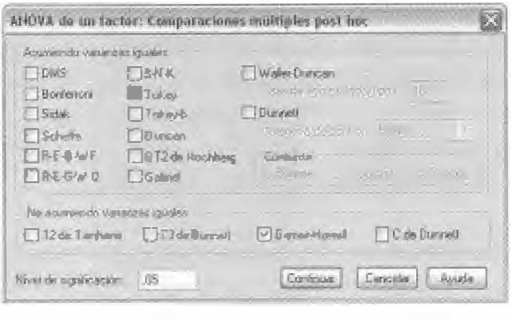
Si pulsamos el botón referido a "Opciones", podremos seleccionar entre una serie de estadísticos básicos que nos permitan operar de mejor forma con este procedimiento. Así, podemos apreciar en la Figura 11 que se nos ofrecen los estadísticos "Descriptivos" típicos (media, desviación típica, error típico de la media, etc.), la posibilidad de establecer los niveles de la variable independiente (en "Efectos fijos y aleatorios"), en caso de que no se quieran utilizar los que normalmente vienen dados por la variable (en tal caso se generan de forma aleatoria), "Prueba de homogeneidad de la varianza" (que ya conocimos como estadístico de Levene), y los estadísticos "Brown-Forsythe" y "Welch" (equivalentes al estadístico F cuando no se pueden asumir varianzas poblacionales iguales). Además, desde este cuadro de diálogo podemos pedir el "Gráfico de las medias" y el tratamiento que se le dará a los valores perdidos.

Figura 11. *ANOVA de un factor: Opciones*



En la opción "Post hoc" del cuadro de diálogo principal encontraremos una serie de opciones que nos permitan especificar, una vez realizado el análisis de varianza y en caso de que existen diferencias significativas entre los grupos, dónde en concreto se encuentran dichas diferencias (Figura 12). Todas las opciones ofrecidas por este cuadro apuntan a lo mismo, esto es a contrastar una vez rechazada la hipótesis de igualdad de medias entre qué grupos se verifican dichas diferencias. Si se asumen varianzas iguales, uno de los métodos con mayor aceptación es el de Tukey. Si no pueden asumirse varianzas iguales, entonces el método que recomendamos (similar al de Tukey) es el de Games-Howell, ya que permite controlar de mejor forma la tasa de error (para una detallada descripción de los métodos ver Pardo y Ruiz, 2002). Además, en este cuadro podemos fijar el nivel de significación con el que queremos trabajar (y cuya opción por defecto es 0.05).

Figura 12. *ANOVA de un factor: Comparaciones post hoc*



Finalmente, la opción "Contrastes" (Figura 13) del cuadro de diálogo para ANOVA de un factor nos permite efectuar comparaciones de tendencia y realizar otro tipo de comparaciones entre medias (contrastes personalizados).

La opción "Polinómico" permite obtener comparaciones de tendencia. Es decir, cuando se rechaza la hipótesis de igualdad de medias entre grupos y contamos con una variable independiente que es de tipo cuantitativo, entonces gracias a esta opción podemos indagar en el tipo de relación que se establece entre las variables dependiente e independiente (lineal, cuadrática, cúbica, etc.). La opción "Orden" asociada a este procedimiento permite especificar cuál es el polinomio de mayor orden que se desea estudiar (se entregan los análisis para todos los niveles inferiores de la opción elegida).

Figura 13. *ANOVA de un factor: Contrastes*

Por otra parte, la opción "Coeficientes" será la que nos permita realizar contrastes personalizados, esto es, asignar coeficientes concretos a cada uno de los grupos que se desea comparar de modo de seleccionar aquellos que se desean mediante la asignación de números que los identifiquen.

Una vez realizadas las especificaciones sobre el modelo de ANOVA volvemos al cuadro de diálogo principal y pulsamos "Aceptar", tras

lo cual encontraremos los resultados obtenidos en el visor habilitado para ello.

El estadístico F se interpreta de la misma forma en que lo hemos venido haciendo hasta ahora, a saber: si el nivel de significación asociado es inferior a 0.05, entonces se rechaza la hipótesis nula de igualdad de medias y se concluye que las medias poblacionales son diferentes. Como podemos apreciar en nuestro ejemplo (Figura 14), el nivel crítico asociado al estadístico F es mayor a 0.05, lo que nos indica que debemos mantener la hipótesis nula de igualdad de medias y concluir que no existen diferencias significativas entre los grupos de diferente nivel socioeconómico, en sus puntajes en la escala. Este resultado haría innecesarios los procedimientos complementarios que hemos especificado (tanto los resultados para la prueba de homogeneidad de varianza, como las comparaciones pos hoc y su respectiva tabla para los subconjuntos homogéneos).

Figura 14. *ANOVA de un factor: Resumen del procedimiento*

ANOVA

PUNTAJE

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	154,618	2	77,309	,987	,375
Intra-grupos	10886,713	139	78,322		
Total	11041,331	141			

Si los resultados hubiesen sido otros (si se hubiese constatado la diferencia de medias), entonces recurriríamos a los resultados de la prueba de homogeneidad de la varianza (estadístico de Levene) que se muestran en la Figura 15 y que nos indica que las varianzas son iguales, o mejor dicho, que no puede rechazarse la hipótesis de igualdad de varianzas, lo que nos indicaría que debemos prestar atención a la solución entregada por el estadístico de Tukey (Figura 16). Allí debemos fijarnos en la significación asociada a la comparación de cada par de grupos, de manera de constatar la diferencia de medias entre estos. La columna de medias nos entrega

adicionalmente información sobre el resultado de dicha comparación, ya que aquellas comparaciones en que la diferencia es significativa aparecen marcadas con un asterisco.

Figura 15. *ANOVA de un factor: Prueba de homogeneidad de la varianza*

Prueba de homogeneidad de varianzas

PUNTAJE

Estadístico de Levene	gl1	gl2	Sig.
,796	2	139	,453

Figura 16. *ANOVA de un factor: Comparaciones post hoc*

Comparaciones múltiples

Variable dependiente: PUNTAJE

	(I) nivel socioeconómico del sujeto	(J) nivel socioeconómico del sujeto	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
HSD de Tukey	alto	medio	1,31	1,882	,767	-3,15	5,77
		bajo	2,86	2,059	,348	-2,01	7,74
	medio	alto	-1,31	1,882	,767	-5,77	3,15
		bajo	1,55	1,742	,646	-2,57	5,68
	bajo	alto	-2,86	2,059	,348	-7,74	2,01
		medio	-1,55	1,742	,646	-5,68	2,57
Games-Howell	alto	medio	1,31	1,975	,786	-3,44	6,06
		bajo	2,86	2,094	,364	-2,16	7,89
	medio	alto	-1,31	1,975	,786	-6,06	3,44
		bajo	1,55	1,678	,625	-2,44	5,55
	bajo	alto	-2,86	2,094	,364	-7,89	2,16
		medio	-1,55	1,678	,625	-5,55	2,44

La tabla de subconjuntos homogéneos (que se genera automáticamente al pedir pruebas post hoc) clasifica los grupos según el grado de semejanza de sus medias (Figura 17) y nos informa del número de personas que pertenecen a cada grupo (N). Para nuestro ejemplo, y debido a que no existen diferencias de medias, se ha clasificado a los tres grupos formados por la variable NSE en un solo conjunto y se ha especificado el nivel de significación.

Figura 17. *ANOVA de un factor: Subconjuntos homogéneos*

PUNTAJE

	nivel socioeconómico del sujeto	N	Subconjunto para alfa = .05
			1
HSD de Tukey [a,b]	bajo	42	19,83
	medio	67	21,39
	alto	33	22,70
	Sig.		,290

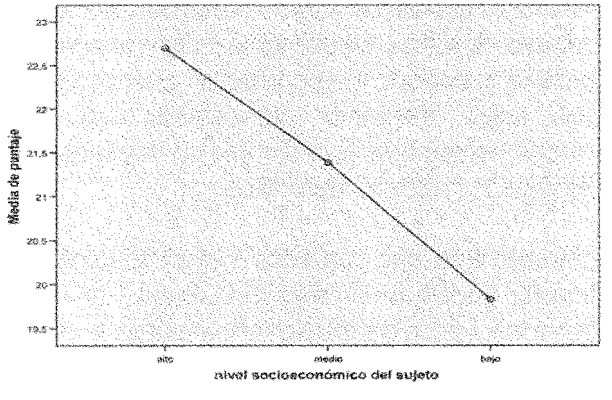
Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 43,454.

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.

Finalmente, hemos pedido al realizar el procedimiento ANOVA de un factor una gráfica de las medias para cada grupo que nos permita apreciar más claramente las diferencias realmente existentes. Así, podemos ver cómo a pesar de no ser significativas las diferencias entre grupos puede detectarse una tendencia a responder levemente más alto a la escala mientras más elevado sea el NSE de los sujetos.

Figura 18. *ANOVA de un factor: Gráfico de las medias*



8.5. Análisis de Varianza (ANOVA) Factorial

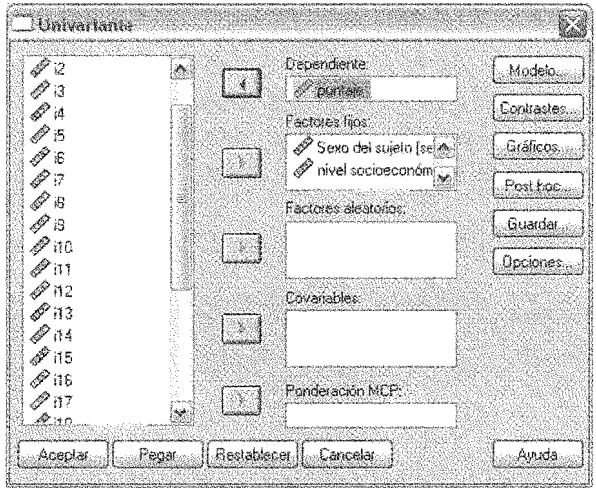
El modelo que pasamos a revisar implica la realización de un análisis de varianza sobre más de un factor (VI). Se utiliza para ver el efecto individual y conjunto (interacción) de dos o más factores (variables independientes categóricas) sobre una variable dependiente cuantitativa. Es decir, que en este modelo, en caso que se consideran solo dos factores, los efectos de interés son tres: los de cada factor principal por separado y el efecto de interacción entre ambos.

Para el caso del análisis de varianza factorial se debe generar una hipótesis nula para cada factor y para cada posible combinación de factores. La hipótesis nula referida a un factor afirma que las medias poblacionales definidas por el factor son iguales. La hipótesis nula referida al efecto de una interacción afirma que tal efecto es nulo. Así, para cada hipótesis se generará un estadístico F (y su respectivo nivel crítico) que permitirá contrastarla y decidir si debe ser rechazada o mantenida la hipótesis.

Para ejecutar este procedimiento se debe pulsar el menú Analizar > Modelo lineal general > Univariante, ante lo cual se desplegará el cuadro de diálogo para dicho análisis (Figura 19). Al lado izquierdo se encuentra el listado de todas las variables contenidas en el archivo de datos. Debemos escoger una variable cuantitativa y trasladarla al cuadro "Dependiente" que aparece en la pantalla. Por otra parte, debemos seleccionar dos o más variables categóricas (nominales u ordinales) y llevarlas a los cuadros de factores fijos o factores aleatorios.

Los factores fijos son aquellos cuyos niveles los establece el investigador (todos los niveles del factor). Los factores aleatorios son aquellos seleccionados de forma aleatoria en todos los niveles posibles del factor (una muestra de niveles del factor).

Figura 19. *Modelo lineal general: Univariante*

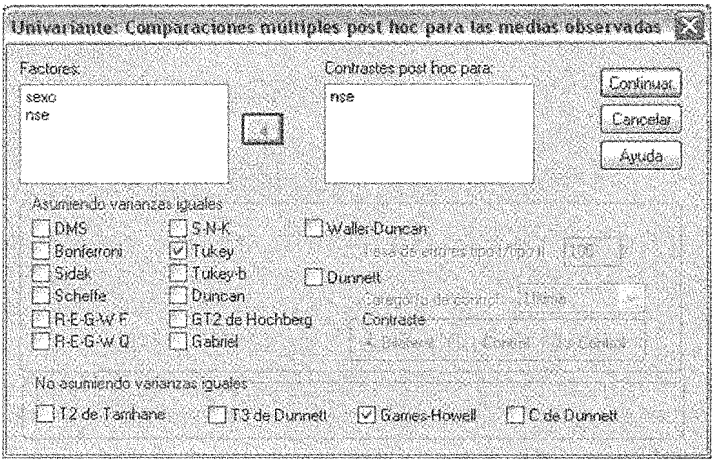


El cuadro de diálogo nos permite, además, realizar análisis de covarianza (permite eliminar de la variable dependiente el efecto atribuible a variables no incluidas en el diseño como factores y no sometidas a control experimental) y utilizar el método de mínimos cuadrados ponderados (MCP), el cual nos permite realizar una estimación óptima cuando las varianzas poblacionales no son iguales.

Para realizar un análisis de covarianza debemos seleccionar una variable cuantitativa y trasladarla al cuadro dependiente del cuadro de diálogo univariante. Por otra parte trasladamos una o más variables categóricas a la lista factores fijos o factores aleatorios. Seleccionamos la variable que deseamos controlar trasladándola al cuadro covariable. En el análisis de covarianza los efectos de interés siguen siendo los referidos a cada factor y a las interacciones entre factores. Este procedimiento también nos permite evaluar el efecto individual de cada una de las covariables incluidas, siendo la hipótesis a contrastar que cada coeficiente de regresión correspondiente a una covariable vale cero en una población. Así, se determina si la covariable posee

o no efecto significativo (está o no relacionada linealmente con la variable dependiente). Si ninguna covariable posee efecto significativo es esperable que los resultados de la ANCOVA sean similares a los de la ANOVA. Si las covariables poseen efectos significativos puede ocurrir que los resultados ANOVA-ANCOVA sean distintos (ya sea porque un efecto no significativo en ANOVA es significativo en ANCOVA o viceversa).

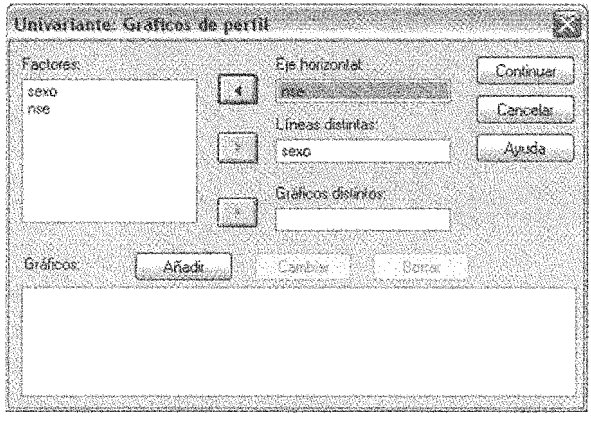
Figura 20. *Modelo lineal general: Comparaciones múltiples*



Si alguno de los efectos principales resulta significativo, puede resultar útil (al igual que en el procedimiento ANOVA de un factor) realizar comparaciones post hoc que nos muestren entre qué grupos se detectan las diferencias de media. Para realizar dichas comparaciones debemos pulsar la opción Post hoc del cuadro de diálogo principal. En el (ver Figura 20) debemos seleccionar las variables sobre las que se realizarán los contrastes y marcar las opciones deseadas para el caso de que se asuman varianzas iguales o que este supuesto no pueda ser mantenido. Los resultados que nos muestra el visor son similares a los arrojados para la ANOVA de un factor, a saber: tabla de comparaciones múltiples y de subgrupos homogéneos.

Para la correcta interpretación de una interacción entre variables podemos generar un gráfico de perfil que nos muestre dicha interacción entre los dos factores. En el eje de ordenadas encontraremos la media para la variable dependiente. En el eje de abscisas se representarán los niveles del primer factor. Las líneas del gráfico representan los niveles del segundo factor. Para generar dicho gráfico pulsamos el botón "Gráficos" del cuadro de diálogo univariante y una vez desplegado el cuadro de diálogo (Figura 21) para dicha opción trasladamos las variables que se desea representar a las ventanas denominadas "Eje horizontal" y "Líneas distintas".

Figura 21. *Modelo lineal general: Gráficos de perfil*



Hasta aquí hemos presentado las opciones más generales que nos permite este procedimiento. Para un análisis detallado de las opciones que entrega SPSS para este procedimiento se puede revisar el manual de Pardo y Ruiz (2002).

Los resultados que nos entrega este procedimiento son básicamente los mismos que los obtenidos para el modelo de ANOVA de un factor, pero contamos con al menos tres estadísticos F, la información sobre interacción de variables y con

la gráfica para dicha interacción. En los casos en que resulte pertinente se realizarán las pruebas post hoc (en nuestro caso esto solo es posible para la variable NSE, ya que posee más de dos niveles). En nuestro caso, la Figura 22 nos muestra la tabla con el resumen del procedimiento de ANOVA factorial, y en ella podemos apreciar que la única variable que tiene incidencia sobre la puntuación es el sexo (sig. < 0.05), ya que ni el NSE ni la interacción Sexo*NSE resultan significativas.

Figura 22. *ANOVA factorial: resumen del modelo*

Pruebas de los efectos inter-sujetos

Variable dependiente: PUNTAJE

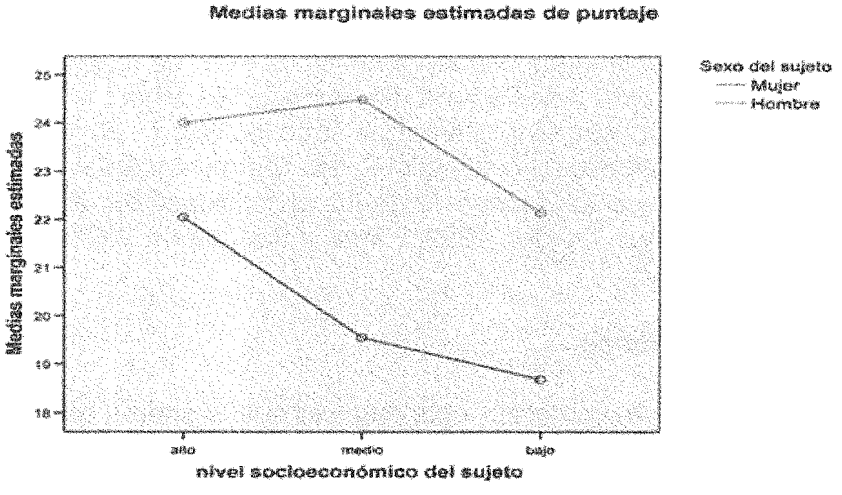
Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	675,910[a]	5	135,182	1,774	,122
Intersección	55751,636	1	55751,636	731,492	,000
SEXO	348,656	1	348,656	4,575	,034
NSE	118,726	2	59,363	,779	,461
SEXO * NSE	46,087	2	23,043	,302	,740
Error	10365,421	136	76,216		
Total	75057,000	142			
Total corregida	11041,331	141			

a. R cuadrado = ,061 (R cuadrado corregida = ,027)

De los resultados obtenidos para la interacción se puede inferir que las diferencias detectadas en la escala para ambos sexos se reproducen de forma similar en los diferentes niveles socioeconómicos. El valor de R cuadrado (que aparece al pie de la tabla) nos indica que las tres variables incluidas en el modelo (Sexo, NSE y Sexo*NSE) explican un 6,1% de la varianza de la variable dependiente Puntaje, en la escala de prejuicio.

Para el caso de los análisis post hoc que se han solicitado, los resultados obtenidos son similares a los observados en el apartado de ANOVA de un factor (ya que solo se pueden realizan para la variable NSE que es la que posee más de dos niveles) y por lo tanto la tabla obtenida es idéntica a la de la figura 16. Lo mismo ocurre para el caso de la tabla de subconjuntos homogéneos.

Figura 23. *ANOVA factorial: gráfico de perfil: NSE*Sexo*



En la figura 23 podemos apreciar el gráfico de perfil en que quedan representadas las medias de las escalas calculadas para cada uno de los grupos resultantes de la combinación de las variables Sexo y NSE. Al fijarnos en la gráfica observamos la falta de interacción de las variables, se observa como cada una sigue un curso propio dependiendo del sexo de los sujetos, sin que se observen claras diferencias por NSE.

Como podemos ver, el modelo de ANOVA factorial nos ofrece la posibilidad de observar el efecto de la interacción de variables sobre nuestra variable dependiente. Obviamente, aquí nos hemos limitado a revisar las operaciones básicas que son posibles de realizar, ya que SPSS nos ofrece otra serie de útiles herramientas y cálculos. Ya hemos mencionado que entre otros, es posible el cálculo de covarianzas (ANCOVA) que permite un adecuado control estadístico de las variables que podrían afectar la variable dependiente. Además, en el cuadro de diálogo Opciones se nos ofrece la posibilidad de comparar los *efectos principales* (es una

comparación por pares de variables similar a la realizada en las pruebas T), observar los *estadísticos descriptivos* (media y desviación típica), *estimación del tamaño del efecto* (grado en que cada factor o combinación de factores está explicando la variación de la variable dependiente), potencia observada (capacidad del contraste para detectar una diferencia poblacional), *gráficos de residuos* (muestra las diferencias entre los valores observados y los pronosticados por el modelo), etc. (para un análisis detallado de los elementos teóricos y de las aplicaciones del análisis de varianza, ver Tejedor, 1999; Pardo y San Martín 2001; Pardo y Ruiz, 2002).

8.6. Correlaciones

Decíamos al comenzar este capítulo que el interés general del análisis de datos gira en torno de la comparación de grupos y del establecimiento de relaciones entre variables. Hasta ahora hemos venido hablando de los procedimientos de comparación, ahora corresponde estudiar lo referido a la relación entre variables cuantitativas, de modo tal que podamos cuantificar el grado de relación existente entre ellas.

El concepto de correlación se refieres al grado de variación conjunta existente entre dos o más variables. El caso de la correlación lineal tiene como límite la relación entre dos variables (correlación simple). Para el caso de más de dos variables, nos referimos al procedimiento regresión lineal.

Una relación lineal positiva implica que dos variables varían de forma parecida. Es decir, los sujetos que puntúan alto en una variable (X) lo hacen también en la otra (Y), y los sujetos que puntúan bajo en la primera variable (X) tienden a puntuar bajo también en la segunda (Y). Por otra parte, una relación lineal negativa significa justamente lo contrario, es decir, los sujetos que puntúan alto en la primera variable (X) logran puntuaciones bajas en la segunda (Y). Inversamente los sujetos que puntúan

bajo en la primera variable (X), logran puntuaciones altas en la segunda (Y).

En este capítulo estudiaremos dos tipos de correlaciones, las bivariadas (se estudia la relación lineal entre dos variables) y las parciales (se estudia la relación lineal entre dos variables, pero controlando o eliminando el efecto atribuible a otras variables).

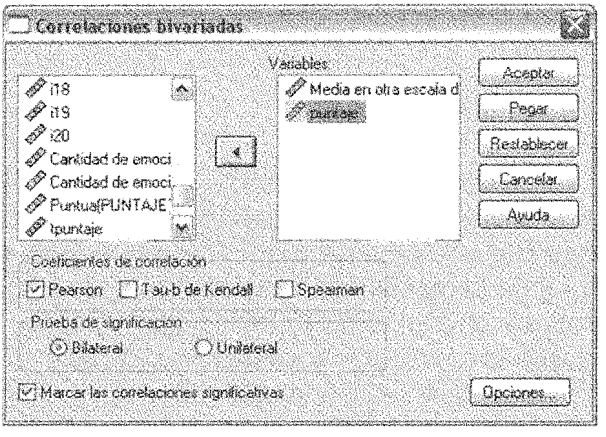
8.6.1. Correlación bivariada

El procedimiento de correlaciones requiere que tanto la variable dependiente como la independiente sean cuantitativas. El supuesto a la base es que el comportamiento de una variable depende del comportamiento de la otra, de modo que podemos predecir el comportamiento de una variable a partir de su relación con otra, aunque ahora sin distinción entre las variables dependiente e independiente. Lo que en concreto hacemos, es definir el sentido de dicha relación y cuantificar la intensidad de la misma.

Es muy importante recordar que la correlación nunca implica causalidad, ya que dos variables pueden estar relacionadas sin que una sea causa de la otra. Es decir, lo que observamos a través de las correlaciones es la ocurrencia conjunta de dos fenómenos (a partir de la cual se infiere relación) y la afectación mutua. Para que quede más claro, de una correlación positiva entre variables como coeficiente intelectual y rendimiento académico no se desprende que una sea la causa de la otra, ya que es perfectamente posible que una persona con un coeficiente intelectual elevado pueda rendir mal debido a otros factores no incorporados en la relación estudiada. En este sentido, con las correlaciones siempre nos mantenemos en el ámbito descriptivo aunque en la práctica las utilicemos para realizar inferencias a partir de la relación lineal entre eventos.

Para realizar un procedimiento de correlaciones en SPSS debemos pulsar el menú Analizar > Correlaciones > Bivariadas, ante lo cual se desplegará el cuadro de diálogo principal de dicho procedimiento (Figura 24).

Figura 24. *Correlaciones: Bivariadas*



El procedimiento Correlaciones bivariadas de SPSS nos ofrece tres coeficientes distintos entre los que podemos escoger:

- rxy de Pearson: se trata del coeficiente más utilizado para estudiar el grado de relación lineal entre dos variables cuantitativas. Toma valores entre -1 (relación lineal perfecta y negativa) y 1 (relación lineal perfecta y positiva). Un valor 0 (cero) implica relación lineal nula.

- tau-b de Kendall: este coeficiente es apropiado para el estudio de variables cuyo nivel de medida es de tipo ordinal. Toma valores entre 1 y -1, los que se interpretan exactamente de la misma forma que en el coeficiente anteriormente revisado.

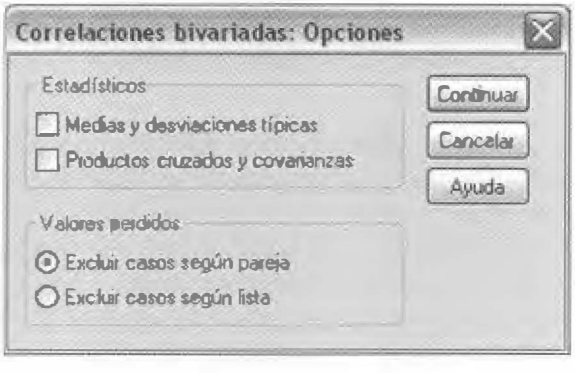
- rho de Spearman: se trata de una operación realizada sobre el coeficiente de correlación de Pearson que transforma las

puntuaciones originales en rangos, Sus valores también fluctúan entre -1 y 1, y se interpreta de forma idéntica a los anteriores. Se utiliza, igual que tau-b, como alternativa a la correlación de Pearson cuando las variables son de tipo ordinal o no se cumple el supuesto de normalidad.

Además de los coeficientes de correlación, el programa nos entregará las respectivas pruebas de significación que nos permitirán contrastar la hipótesis nula de que el valor poblacional del coeficiente es cero (es decir, que no habría relación entre las variables). El rechazo de la hipótesis de independencia lineal nos permitirá afirmar que las variables están relacionadas de forma significativa.

Por otra parte, el cuadro de diálogo nos ofrece la posibilidad de escoger entre las opciones Bilateral y Unilateral. Cuando no existen expectativas de la dirección de la relación entre variables debemos ocupar la opción Bilateral. Para el caso en que dichas expectativas sobre la dirección de la relación si existan, entonces la opción más útil es Unilateral.

Figura 25. *Correlaciones bivariadas: Opciones*



El cuadro de diálogo anterior nos ofrece también ciertas opciones que nos entregarán información adicional pertinente para el proceso que estamos analizando (estadísticos

descriptivos, covarianzas, etc.), así como definir el tratamiento que queremos dar a los valores perdidos. La figura 25 nos muestra el subcuadro de diálogos opciones. En dicho cuadro podemos pedir las medias y desviaciones típicas para las variables utilizadas, así como la covarianza (los productos de las desviaciones de cada puntuación respecto de su media). Por otra parte, podemos decidir qué hacer con los valores perdidos, ya sea excluirlos según pareja (se excluyen del cálculo de correlación los casos con valor perdido en alguna de las dos variables) o excluirlos según lista (se excluyen los casos con valor pedido en cualquiera de las variables de la lista de variables). Una vez seleccionadas las opciones pulsamos aceptar y revisamos las tablas ofrecidas en el visor de resultados.

Figura 26. *Correlación de Pearson*

Correlaciones

		Media en otra escala de homofobia	PUNTAJE
Media en otra escala de homofobia	Correlación de Pearson	1	,818**
	Sig. (bilateral)	.	,000
	N	142	142
PUNTAJE	Correlación de Pearson	,818**	1
	Sig. (bilateral)	,000	.
	N	142	142

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

En la tabla ofrecida por SPSS para el coeficiente de correlación de Pearson podemos observar (ver Figura 26) que cada celda contiene los valores referidos a la correlación entre las dos variables, a la significación y al número de sujetos de la muestra que se consideran casos válidos para el análisis. Para el caso de nuestro ejemplo, el nivel crítico (significación) asociado nos permite rechazar la hipótesis nula que refiere a que el coeficiente de correlación vale cero y, por lo tanto, debemos concluir que existe una relación lineal positiva que es significativa. De este modo, entre las variables PUNTAJE

(que sabemos que corresponde a la media de puntaje en la escala de prejuicio) y Media en otra escala (puntaje en la escala de homofobia), encontramos una correlación positiva y alta (.81), toda vez que significativa (sig. = 0,000). También observamos que se ha trabajado sobre una muestra de 142 casos válidos. Ya hemos indicado que es posible obtener otros coeficientes de correlación (Kendall y Spearman). En la Figura 27 se pueden observar los resultados para dichos procedimientos. Como ya hemos indicado, estos coeficientes son apropiados para el caso de variables medidas a nivel ordinal, por lo que los resultados obviamente no son coincidentes con los del coeficiente de correlación de Pearson y solo deben considerarse a título ilustrativo.

Figura 27. *Correlaciones de Kendall y Spearman*

Correlaciones

			Media en otra escala de homofobia	PUNTAJE
Tau_b de Kendall	Media en otra escala de homofobia	Coeficiente de correlación	1,000	,599**
		Sig. (bilateral)	.	,000
		N	142	142
	PUNTAJE	Coeficiente de correlación	,599**	1,000
		Sig. (bilateral)	,000	.
		N	142	142
Rho de Spearman	Media en otra escala de homofobia	Coeficiente de correlación	1,000	,769**
		Sig. (bilateral)	.	,000
		N	142	142
	PUNTAJE	Coeficiente de correlación	,769**	1,000
		Sig. (bilateral)	,000	.
		N	142	142

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

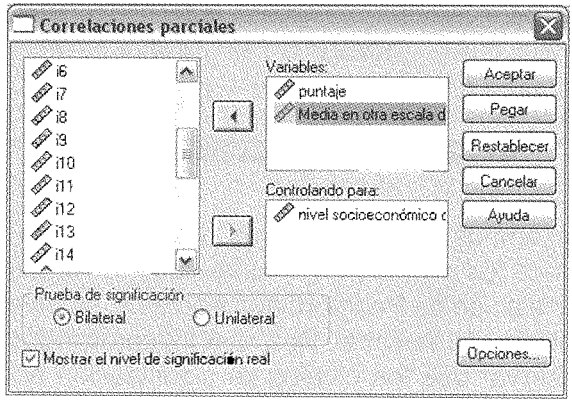
En ambos casos podemos observar cómo cada celda nos ofrece los mismos tres valores que la tabla de coeficiente de correlación de Pearson: el coeficiente, su nivel de significación y el tamaño muestral. Ya a nivel interpretativo, también podemos observar que los niveles de significación nos autorizan para

rechazar la hipótesis nula que afirmaba la independencia de las variables, aunque con coeficientes de correlación inferiores.

8.6.2. Correlación Parcial

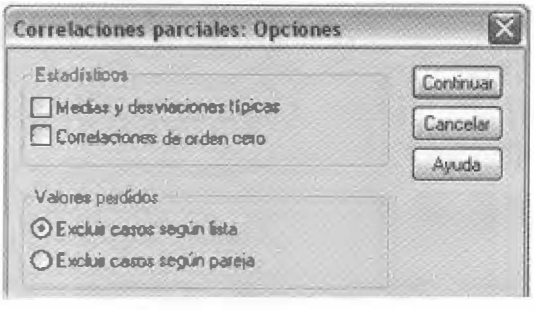
Como ya indicábamos al comenzar este apartado, este procedimiento permite estudiar la relación lineal entre dos variables controlando el efecto de una o más variables extrañas. Es decir, se intenta eliminar el efecto atribuible a terceras variables que pudieran condicionar o modular la relación entre las variables estudiadas. Para obtener el coeficiente de correlación parcial debemos pulsar la opción correlaciones parciales del menú analizar y se desplegará el cuadro de diálogo correspondiente (Figura 28). La única innovación que debemos hacer respecto del procedimiento de correlación bivariada es agregar las variables que se desean controlar traspasándolas a la ventana correspondiente. La opción Mostrar el nivel de significación real viene dada por defecto y permite observar los niveles críticos exactos asociados al coeficiente de correlación. En caso de desactivarlo aparecerá un asterisco en aquellos casos en que la significación sea menor a 0,05 y dos asteriscos cuando el nivel crítico sea menor a 0,01.

Figura 28. *Correlaciones parciales*



Si pulsamos el cuadro de diálogo de "Opciones" (Figura 29), podremos obtener información adicional sobre los estadísticos descriptivos (media y desviación típica), así como de las correlaciones de orden cero (se entregan las correlaciones para cada par de variables sin ejercer control sobre las variables controladas). Además, y como en todos los procedimientos anteriores, podemos controlar las opciones referidas al tratamiento de los datos perdidos (excluir casos según lista o según pareja).

Figura 29. *Correlaciones parciales: Opciones*



La Figura 30 nos muestra los resultados obtenidos mediante el procedimiento de correlaciones parciales, utilizando un coeficiente de correlación de Pearson. Como vemos, el control de la variable NSE nos indica que esta no tiene efecto significativo sobre la correlación entre las variables PUNTAJE y MED_HATH (media en escala de homofobia), no alterando la variable NSE sustancialmente la relación que se establece entre nuestras variables principales.

Figura 30. *Correlaciones parciales*

Correlaciones

Variables de control			puntaje	Media en otra escala de homofobia
nivel socioeconómico del sujeto	puntaje	Correlación	1,000	,818
		Significación (bilateral)	.	,000
		gl	0	139
	Media en otra escala de homofobia	Correlación	,818	1,000
		Significación (bilateral)	,000	.
		gl	139	0

La matriz ofrece para el par de variables seleccionado el coeficiente de correlación de Pearson (.81), los grados de libertad asociados (139) y el nivel crítico (0,000). De este modo, podemos concluir que existe una correlación lineal positiva y alta entre las variables estudiadas y que dicho efecto no se debe a la variable controlada (NSE) o no está mediado por dicha variable.

Cabe destacar que el modo en que se ha utilizado el procedimiento de correlaciones en nuestro ejemplo es útil como indicación de validez del instrumento, ya que como hipótesis de partida era esperable que una persona que manifiesta alto prejuicio en una escala, también lo hiciera en la otra. Es decir, ambas escalas parecen estar midiendo el mismo constructo.

8.7. Chi Cuadrado (X^2)

El procedimiento que ahora analizamos, a diferencia de los anteriores, tiene la particularidad de trabajar con lo que denominamos datos cualitativos. Este tipo de datos refiere a aquellas variables que han sido medidas a nivel nominal (o también denominadas categóricas). Se trata de todos aquellas medidas que nos permiten clasificar de forma exhaustiva y exclusiva a los sujetos de una muestra, pero permitiéndonos tan solo establecer relaciones de igualdad o diferencia en alguna característica de interés. De un modo más claro, el procedimiento X2 se utiliza cuando contamos con dos variables de tipo cualitativo. Por ejemplo, si contamos con las variables "Sexo" (0= hombre y 1= mujeres) y "Categoría laboral" (1= directivo, 2= administrativo y 3= operario) y queremos establecer si existe relación entre ser hombre o mujer y pertenecer a una determinada categoría laboral, entonces el procedimiento adecuado es Chi-cuadrado.

Las variables cualitativas son muy frecuentes en psicología, y aunque existen diversas técnicas, dependiendo del número

de variables (una dos o más), el tipo de diseño (transversal o longitudinal) y el tipo de variables (dicotómicas o politómicas), nosotros solo nos centraremos en aquellas que se ajustan al procedimiento revisado. Para una revisión exhaustiva de otros procedimientos para el estudio de tablas de contingencia multidimensionales (contraste de hipótesis sobre proporciones, modelos log-lineales, etc.), se puede revisar el manual de análisis de datos de Pardo y San Martín (2004) y de Pardo y Ruiz (2002) o, en general, cualquier manual sobre análisis de datos.

Por otra parte, el procedimiento X2, permite estudiar diferentes aspectos del análisis referido a variables cualitativas (bondad de ajuste, homogeneidad, igualdad de proporciones, etc.), pero aquí se estudiarán solo los aspectos referidos al contraste de la independencia entre dos variables. Es decir, contrastaremos la hipótesis nula de que las dos variables (por ejemplo Sexo y Categoría laboral) son independientes. Si rechazamos dicha hipótesis, deberemos afirmar que las variables están relacionas. Es decir, con este procedimiento nos quedamos siempre en el plano descriptivo, ya que lo único que podemos hacer es detectar pautas de asociación entre variables.

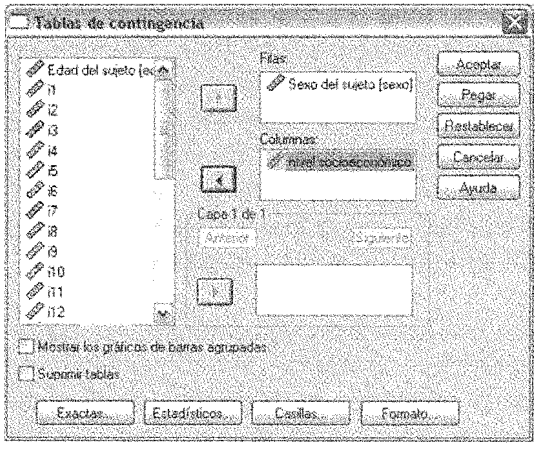
Al trabajar con variables categóricas los datos son manejados en tablas de doble entrada, en las que cada entrada representa una variable diferente, de modo que en cada casilla se expresa la frecuencia o los porcentajes de aparición del cruce de ambas entradas. A estas tablas de doble entrada se les llama tablas de contingencia. El número de entradas depende de la cantidad de variables de las que disponemos y SPSS permite realizar tablas con cuantas queramos, aunque los estadísticos útiles para su análisis solo sirven para tablas bidimensionales (aquella que cruza dos variables).

Ahora bien, lo que sí es posible es utilizar terceras variables a modo de segmentación, de modo tal de dividir la muestra

en subgrupos o capas. El procedimiento Tablas de contingencia contiene una serie de estadísticos y medidas de asociación que proporcionan la información necesaria para indagar en las pautas de asociación de las diferentes variables de nuestra tabla.

Para ejecutar el procedimiento Tablas de contingencia, seleccionamos dicha opción desde Estadísticos descriptivos del menú Analizar. Al realizar dicho procedimiento se desplegará el cuadro de diálogo de dicha opción (figura 31).

Figura 31. *Tablas de contingencia*



Como podemos observar, el cuadro de diálogo nos ofrece a un costado las variables que contiene nuestro archivo y tan solo debemos traspasar una variable categórica a la lista Filas y la otra a la que pide especificar las Columnas. Marcando la opción "Mostrar los gráficos de barras agrupadas" podremos obtener la tabla de frecuencias para las variables que hemos cruzado, de modo de poder observar el gráfico que nos resume dichas variables (y en el cual cada barra corresponderá a una casilla de la tabla).

En caso de marcar la opción "Suprimir tablas", entonces el visor de resultados solo nos entregará el gráfico de barras agrupadas. Si revisamos los estadísticos disponibles para ejecutarse desde la opción del mismo nombre veremos (Figura 32) una serie de opciones, entre las que encontramos nuestra medida de asociación (y una serie de medidas para variables nominales y ordinales).

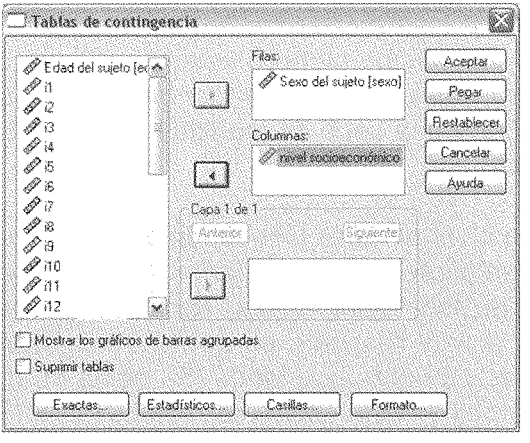
Para que el estadístico Chi-cuadrado sea una buena elección deben cumplirse algunos criterios básicos tales como que las frecuencias esperadas no sean demasiado pequeñas. El criterio utilizado a este respecto es que cuando existen frecuencias esperadas menores que cinco (5), estas no deben superar el 20% del total de frecuencias esperadas (en SPSS se nos ofrece una nota al pie de la tabla respectiva indicándonos el valor de la frecuencia esperada más pequeña, así como de los casos en que esta es menor a 5 y su respectivo porcentaje).

En caso de que el porcentaje supere el 20%, Chi-cuadrado podría no ser una buena idea o al menos deberá ser interpretado con suma cautela.

Como podemos apreciar en la Figura 32, para el caso de datos nominales es posible solicitar una serie de estadísticos que nos permitan cuantificar la fuerza de la asociación entre las variables (ya que como habíamos precisado con anterioridad, X2 nos ofrece simplemente información –gracias al contraste de hipótesis- respecto de si existe o no asociación entre las variables e intentando eliminar el efecto que el tamaño de la muestra tiene sobre el valor de X2). Eso sí, a diferencia del procedimiento de correlaciones, aquí no tiene mucho sentido hablar de la dirección de la asociación.

Es por ello que todos los valores de los coeficientes de asociación se entregan entre 0 y 1 (siendo 0 ausencia de asociación y 1 máxima asociación o asociación perfecta).

Figura 32. *Tablas de contingencia: Estadísticos*



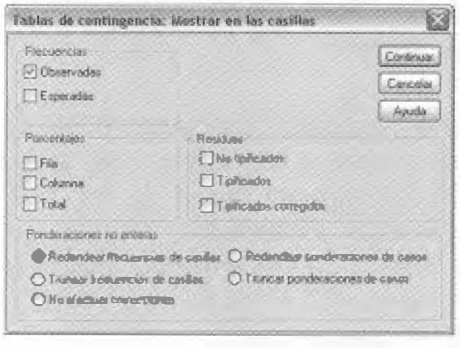
Las medidas disponibles son cuatro: Coeficiente de contingencia, Phi y V de Cramer, Lamba y Coeficiente de incertidumbre. Las dos primeras son medidas basadas en Chi-cuadrado y las dos últimas intentan una reducción proporcional del error (intentan reducir la posibilidad de cometer un error de predicción cuando en lugar de utilizar un caso o grupo de casos pertenecientes a una categoría de la variable, se les clasifica teniendo en cuenta las probabilidades de las categorías de esa variable en cada categoría de una segunda variable).

El cuadro de diálogo principal nos ofrece además algunas otras opciones interesantes, las que encontramos al pulsar el botón "Casillas" (Figura 33), que es el que nos permitirá determinar el contenido de las casillas de la tabla de contingencia y con ello realizar una apropiada interpretación de las pautas de asociación presentes en una tabla después de que algún estadístico conduce al rechazo de la hipótesis de independencia.

La opción referida a las frecuencias nos permitirá apreciar las frecuencias "Observadas" (número de los casos resultantes de la clasificación) y las "Esperadas" (número de casos que debería

haber en cada casilla si las variables fueran independientes). Como vemos, en conjunto estas frecuencias nos permitirán un examen más acucioso de los datos que permita obtener información extra cuando testeamos la hipótesis de independencia entre variables. Por otra parte, la opción "Porcentajes" nos ofrece la posibilidad de apreciar los porcentajes que la frecuencia observada de una casilla representa respecto del total marginal de su fila ("Fila"), el porcentaje que la frecuencia observada de una casilla representa respecto del total marginal de su columna ("Columna") y la frecuencia observada de una casilla representa respecto del número total de casos de la tabla ("Tabla").

Figura 33. *Tablas de contingencia: Casillas*



Finalmente, podemos pedir información referida tanto a los "Residuos" "No tipificados" (diferencia entre la frecuencia esperada y la observada) como "Tipificados" (indicadores de grado en que cada casilla contribuye al valor del estadístico Chi-cuadrado. Si se suman los cuadrados de los residuos tipificados se obtiene el valor del estadístico Chi-cuadrado) y "Tipificados corregidos" (son fácilmente interpretables ya que usando un nivel de confianza del 95% puede afirmarse que los residuos mayores a 1.96 delatan que las casillas correspondientes poseen más datos de los que deberían haber si las variables fueran independientes. Los residuos menores a -1.96 delatan casillas con menos casos que los que cabría esperar bajo la condición de independencia).

En tablas de contingencia que contienen variables nominales, una vez establecida la asociación significativa entre variables (mediante Chi-cuadrado) y que se ha cuantificado la asociación con algún índice de asociación, los residuos tipificados constituyen una poderosa herramienta para interpretar con precisión el significado de la asociación detectada.

En la Figura 34, podemos apreciar los resultados obtenidos al ejecutar los procedimientos que venimos describiendo. La tabla nos muestra el valor del estadístico (valor= .246), sus grados de libertad (gl= 2) y la significación asociada (p>.05). Dichos datos nos indicarían que la hipótesis de independencia entre las variables Sexo y NSE es corroborada, o más bien que nuestros datos son compatibles con dicha hipótesis de independencia. Es decir, no existiría relación entre ser hombre o mujer y pertenecer a un determinado NSE.

Por otra parte, podemos apreciar en el pie de página de la tabla que no existen casillas con frecuencia esperada menor a 5, por lo cual podemos suponer que X^2 es una buena opción para los datos que tenemos. Finalmente, la tabla nos ofrece el valor del estadístico "Razón de verosimilitud" cuyo valor se interpreta de la misma forma que X2 y que es muy útil para estudiar la relación entre variables categóricas.

Figura 34. *Pruebas de Chi-cuadrado*

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	,246[a]	2	,884
Razón de verosimilitud	,246	2	,884
Asociación lineal por lineal	,002	1	,967
N de casos válidos	142		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 11,62.

En la Figura 35 podemos apreciar la cuantificación realizada para el índice de asociación (en este caso resulta poco útil ya que hemos mantenido la hipótesis de independencia de las variables). Observamos que los valores (y su respectiva significación) son similares para las tres medidas y en todos ellos los niveles de asociación son sumamente bajos (.042) y cercanos a cero, lo que corroboraría la hipótesis de independencia de las variables.

Figura 35. *Medidas de asociación simétricas*

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,042	,884
	V de Cramer	,042	,884
	Coeficiente de contingencia	,042	,884
N de casos válidos		142	

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Finalmente, la Figura 36 nos muestra la tabla de contingencia para las variables que venimos utilizando y su respectiva información respecto de los "Residuos no tipificados", "Residuos tipificados" y "Residuos corregidos". El primero de estos nos informa la diferencia entre la frecuencia observada ("Recuento") y esperada. Los residuos corregidos nos indican que nuestras casillas no poseen ni más ni menos datos de los que deberían haber (dado el caso de que nuestras variables son independientes), ya que todos los valores se encuentran dentro del rango que va entre 1.96 y -1.96 (95% de confianza).

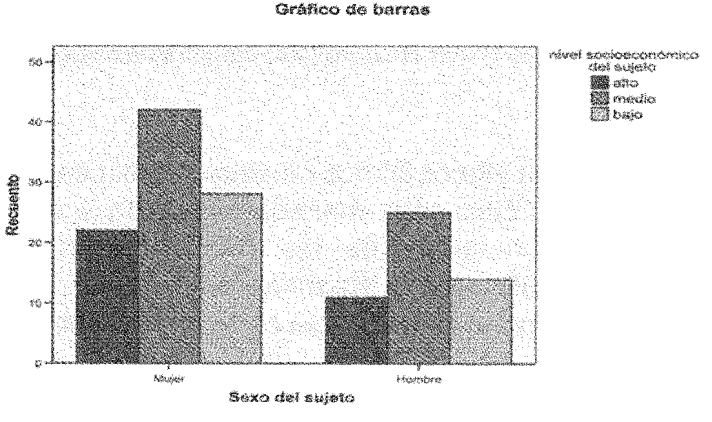
Figura 36. *Tabla de contingencia: Casillas*

Tabla de contingencia Sexo del sujeto * nivel socioeconómico del sujeto

			nivel socioeconómico del sujeto			
			alto	medio	bajo	Total
Sexo del sujeto	Mujer	Recuento	22	42	28	92
		Frecuencia esperada	21,4	43,4	27,2	92,0
		Residuo	,6	-1,4	,8	
		Residuos tipificados	,1	-,2	,2	
		Residuos corregidos	,3	-,5	,3	
	Hombre	Recuento	11	25	14	50
		Frecuencia esperada	11,6	23,6	14,8	50,0
		Residuo	-,6	1,4	-,8	
		Residuos tipificados	-,2	,3	-,2	
		Residuos corregidos	-,3	,5	-,3	
Total		Recuento	33	67	42	142
		Frecuencia esperada	33,0	67,0	42,0	142,0

Finalmente, en La Figura 37 podemos apreciar graficadas las frecuencias que la tabla de contingencias arroja para el cruce de las variables Sexo y NSE.

Figura 37. *Gráficos: Sexo*NSE*



En fin, como podemos apreciar, se trata de un procedimiento muy simple y útil para la detección de pautas de relación entre variables cuando estas son de tipo cualitativo.

8.8. Métodos Multivariados

En términos generales podemos afirmar que reciben el nombre de métodos multivariados (también denominados multivariantes) aquellos que permiten analizar simultáneamente conjuntos amplios de variables interrelacionadas. Estas variables múltiples pueden ser consideradas como dependientes o independientes y correlacionadas entre sí en grados diversos, toda vez que se suponen con distribución normal (Martínez Arias, 1999). De este modo, el análisis multivariante (AM) es aquel cuyos métodos analizan las relaciones entre un número amplio de medidas tomadas sobre cada objeto o unidad de análisis.

En el AM, todas las variables deben ser aleatorias y estar interrelacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido. El propósito de este tipo de análisis es medir, explicar y predecir el grado de relación de los valores teóricos (combinaciones ponderadas de variables), por lo que el carácter multivariante no reside solamente en el número de variables utilizadas (Hair, Anderson, Tatham y Black, 2000).

El AM es utilizado para una serie diversa de procesos entre los que destaca la reducción de datos (se intenta simplificar la estructura del fenómeno estudiado, proporcionándole la estructura más simple posible que facilite su interpretación), clasificación y agrupación (se crean grupos de variables similares entre sí a partir de las características medidas), análisis de las relaciones de dependencia (con propósitos de predicción o explicación) y construcción de modelos y pruebas de hipótesis (algunas técnicas son de carácter descriptivo, pero otras pueden poner a prueba hipótesis sobre complejos basados en poblaciones multivariantes) (Martínez Arias, 1999).

En el contexto de AM una variable es una magnitud de respuesta que representa alguna característica de los objetos y cuyos valores

son el objeto de estudio de la investigación. En nuestro caso, y como veremos más adelante, hemos utilizado matrices de distancias, las que nos entregan medidas de similitud que reflejan la distancia existente entre dos puntos.

Las técnicas de AM se clasifican en, según sea la cantidad de variables dependientes y según cual sea su nivel de medida, métodos de independencia y métodos de dependencia. A continuación nos referiremos brevemente a estos dos grupos.

- *Métodos de Dependencia.* En los denominados métodos de dependencia el interés se centra en la asociación de conjuntos diferenciados de variables y se intenta determinar el grado de relación existente entre los dos conjuntos de variables a partir de un conjunto de variables predictoras.

  Entre las principales técnicas de dependencia encontramos la regresión lineal múltiple, la correlación lineal múltiple, el análisis discriminante, la regresión logística, el análisis multivariante de la varianza y la covarianza (MANOVA y MANCOVA), el análisis de correlación canónica y los modelos de ecuaciones estructurales. Estos métodos no los abordaremos debido a que por los tipos de datos recopilados y los fines de nuestra investigación, solo utilizaremos técnicas de independencia. En todo caso para un buen resumen de sus aplicaciones hay varios textos disponibles (Martínez Arias, 1999; Hair, Anderson, Tatham y Black, 2000).

- *Métodos de Independencia.* Las técnicas de independencia se centran en la relación mutua entre todas las variables y su intención es encontrar información sobre la estructura subyacente o latente a un conjunto de datos, simplificando las complejidades originales, por medio de la reducción de datos, y sin distinción entre variables dependientes e independientes (Martínez Arias, 1999).

Entre las principales técnicas de independencia encontramos el análisis de componentes principales, el análisis factorial, el escalamiento multidimensional, el análisis de conglomerados, los modelos log-lineales (para tablas de frecuencias) y el análisis de correspondencias. La mayor parte de ellas las hemos de utilizar para nuestros fines y las describiremos separadamente, aunque de forma necesariamente breve, en los parágrafos siguientes.

- *Tipos de Datos.* Por otra parte, existen dos tipos de datos con los que se puede trabajar: no métricos (cualitativos) y métricos (cuantitativos). En general, se puede decir que los datos no métricos son atributos, características o propiedades categóricas que identifican o describen a un objeto. Describen diferencias de tipo y clase indicando la presencia o ausencia de una característica o propiedad. Los datos cualitativos pueden presentarse como escalas nominales u ordinales. Las medidas de datos métricos indicarían diferencias de grado o magnitud entre los objetos o individuos, y pueden ser presentadas en escala de intervalo o razón.

8.8.1. Análisis Factorial Exploratorio

El análisis factorial es una técnica de reducción de la dimensionalidad de los datos con la que se pretende encontrar factores comunes que expliquen la presencia de correlaciones entre las variables (Martínez Arias, 1999; Pardo y Ruiz, 2002). De este modo, el análisis factorial (AF) es una invitación a encontrar las variables fundamentales que intervienen en la explicación de ciertos fenómenos, apoyada en la búsqueda de características que presiden las relaciones matemáticas que se establecen a partir de un conjunto de datos (García Jiménez, Gil Flores y Rodríguez Gómez, 2000). De este modo, podríamos afirmar que en el análisis factorial lo que se trata de dilucidar son los elementos definitorios comunes a un grupo

de variables y los componentes específicos que lo caracterizan. O de otro modo, el AF es útil para analizar interrelaciones entre un gran número de variables y explicar estas variables en términos de sus dimensiones subyacentes comunes (factores). El objetivo es encontrar un modo de condensar la información contenida en el conjunto de variables originales en un número más pequeño de variables (factores) con una pérdida mínima de información, procurando que unos grupos de variables sean independientes de otros, pero donde los elementos que componen cada factor tengan con alta correlación entre sí (Hair, Anderson, Tatham y Black, 2000; Pardo y Ruiz, 2002).

Lo que se busca por medio de este proceso es conocer cuál es el mínimo número de factores comunes y distintos necesarios para explicar las correlaciones obtenidas, esto es, condensar y resumir la información sobre variables en un número pequeño de factores. Se busca obtener las dimensiones de variabilidad común a un determinado campo de fenómenos que se ha operativizado por medio de ciertas variables. De aquí la importancia de la elección de variables, ya que no hay posibilidad de que se ponga de manifiesto un determinado factor, si no existe una variable capaz de saturarlo. La elección de variables debe hacer emerger los factores esperados por la teoría que orienta nuestras indagaciones. De este modo, el primer paso del AF es definir el dominio a investigar y la estructura factorial hipotética para dicho dominio (García Jiménez, Gil Flores y Rodríguez Gómez, 2000). Se suele recurrir a variables factorialmente complejas, esto es, representadas por más de una variable con peso alto. Todas las variables del análisis factorial tienen el mismo rango (no existe variable dependiente), esto es, son consideradas como independientes en el sentido de que no existe a priori una dependencia conceptual de una variables sobre otras (Pardo y Ruiz, 2002).

Cuando queremos acercarnos a entidades desconocidas a partir del conocimiento de otras manifiestas, le llamamos análisis

factorial exploratorio. Si queremos ir más allá y explicar las variables en términos de dependencia e independencia, entonces estamos frente al denominado análisis factorial explicativo (García Jiménez, Gil Flores y Rodríguez Gómez, 2000). De un modo más específico, lo que se pretende con el AF es identificar una estructura mediante el resumen de datos, o, simplemente, reducir datos (Hair, Anderson, Tatham y Black, 2000).

Respecto del muestreo podemos afirmar que a mayor tamaño de la muestra más nos acercaremos a las verdaderas puntuaciones del conjunto de la población, pues representa un grado de variabilidad alto en las respuestas. En general, se apunta a que el número de variables no debe exceder a la mitad de los sujetos de la muestra. Las correlaciones obtenidas nos entregan una información inicial sobre el grado de variabilidad observada en las variables del estudio. La varianza explicada dependerá a su vez del método de extracción utilizado. Entre los métodos más utilizados está el de componentes principales (que supone maximizar la varianza explicada al conseguir que la contribución del factor a algunas de las comunalidades de las variables sea máxima.

De allí que el factor que mejor explique la dimensión analizada se convierta en el primer componente principal y así sucesivamente), máxima verosimilitud (considera la mejor estimación posible de la matriz de correlaciones reproducida en la población como principio de extracción, esto es, encontrar la solución factorial que mejor se ajusta a las correlaciones observadas) y alfa (maximiza la generalizabilidad de los factores como principio de extracción, de modo que los factores extraídos tienen correlaciones máximas con el universo de los factores comunes existentes). La comunalidad representa la proporción de la varianza con la que contribuye cada variable a la solución final (Hair, Anderson, Tatham y Black, 2000). El objetivo principal de la fase de extracción es determinar el número mínimo de

factores comunes capaces de reproducir las correlaciones entre las variables observadas

Para hacer más fácil la interpretación de los valores que presenta la matriz factorial tras la extracción, se puede realizar un procedimiento de rotación. De este modo se podrá buscar una mayor simplicidad en los factores (quartimax), en las variables (varimax) o en ambas a la vez (equamax). A este tipo de rotaciones se les denomina ortogonales (pues los ejes se mantienen en ángulos rectos).

En términos generales, podemos afirmar que cuando las correlaciones entre variables son bajas, probablemente no comparten factores comunes. Previamente a los procesos de extracción y rotación debemos analizar las condiciones de aplicación y los índices de adecuación muestral (test de esfericidad de Bartlett, medida de adecuación de muestreo KMO y análisis de residuales), de modo de corroborar el adecuado ajuste de los datos y verificar que el AF es una buena estrategia.

En el proceso de interpretación de los datos se avanza en el sentido de etiquetar cada uno de los factores con varianza alta y que contienen los ítems sobre los que inicialmente se ejecutó el AF, asignándole un nombre que permita agrupar dichas variables comunes.

Para comenzar el análisis en SPSS se debe pulsar Reducción de datos del menú analizar y escoger la opción Análisis factorial. En la Figura 39 podemos apreciar el cuadro de diálogo principal de dicho procedimiento La lista de variables contiene todas aquellas dimensiones contenidas en el archivo, incluidas las de cadena. De este modo debemos transportar aquellas variables que deseamos analizar y mantener, en el lado izquierdo de la pantalla, aquellas que no nos interesen en el análisis.

Figura 39. *Análisis Factorial*

Se puede observar que en el cuadro principal hay una serie de submenús (descriptivos, extracción, puntuaciones, rotación y opciones), cada uno de los cuales nos permitirá definir adecuadamente el modelo de análisis a realizar.

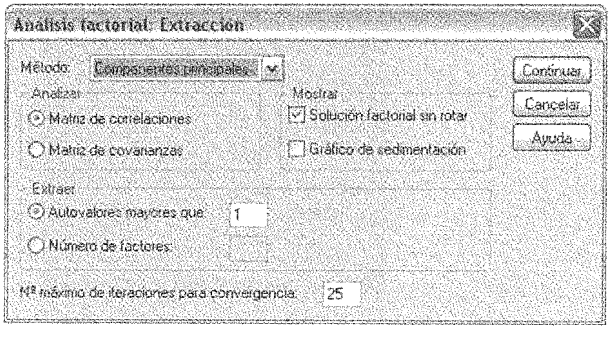
El menú "Descriptivos" (Figura 40) nos permite disponer de los descriptivos univariados más comunes, tales como la media, la desviación típica y el número de casos válidos para el análisis. Toda esta información, como se verá posteriormente, es altamente relevante a la hora de analizar los resultados, debido a que nos informará sobre el grado de dispersión de los puntajes de nuestra muestra y nos permitirá visualizar estos elementos para una fácil inspección. Por otra parte, nos permite observar la solución factorial inicial y su correspondiente comunalidad.

Figura 40. *Análisis Factorial: Descriptivos*

En general, podemos afirmar que en este menú se encontrarán aquellos elementos que nos permitirán contrastar nuestros datos y revisar su adecuación. Es decir, mediante ellos podremos decidir si nuestros datos se ajustan de modo adecuado y si por ello resulta pertinente realizar un análisis factorial. Este cuadro es el que nos entrega la información necesaria para validar los datos y en él podemos proceder a la selección de aquellos aspectos concernientes a la matriz de correlaciones (coeficientes, niveles de significación, determinante, KMO y prueba de esfericidad de Bartlett, inversa, reproducida y anti-imagen).

Al ingresar al cuadro de diálogo Extracción (Figura 41), podemos seleccionar las distintas opciones referidas al método de extracción (componentes principales, mínimos cuadrados no ponderados, mínimos cuadrados generalizados, máxima verosimilitud, factorización de ejes principales, factorización alfa y factorización imagen), del procedimiento de análisis (matriz de covarianzas o de correlaciones), del criterio para determinar en número de factores a extraer (ya sea eligiendo el número de factores o el autovalor más bajo para estos), la solución factorial sin rotar y el gráfico de sedimentación, así como el número máximo de iteraciones deseado para lograr la convergencia.

Figura 41. *Análisis Factorial: Extracción*



El menú Rotación (Figura 42) nos permite escoger el método de rotación ortogonal deseado (varimax, oblimin, quartimax, equamax y promax). En caso de que el ajuste de los datos sea bueno y el análisis no requiera una simplificación, entonces no se justificaría la rotación. Podemos, por otra parte, pedir que se nos muestre la solución rotada y el gráfico de saturaciones para la solución rotada. Finalmente, podemos determinar el número máximo de iteraciones deseado para la convergencia de los datos rotados. De este modo, seleccionamos nuestras opciones entre las posibilidades ofrecidas por el programa y pulsamos el botón continuar, lo que nos devolverá nuevamente al cuadro de diálogo principal y desde donde podremos acceder a los menús restantes.

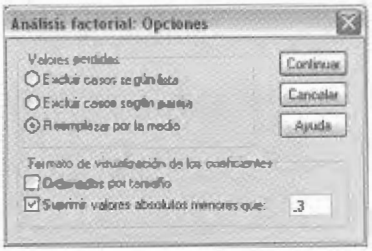
Figura 42. *Análisis Factorial: Rotación*

El menú correspondiente a Puntuaciones nos ofrece la posibilidad de obtener puntuaciones factoriales que proyecten a cada individuo de la muestra sobre cada uno de los factores extraídos del análisis.

Finalmente, el menú opciones (Figura 43) despliega toda una serie de alternativas altamente interesantes para complementar nuestro análisis. En principio debemos tomar una decisión sobre qué hacer con los valores perdidos, esto es, sobre aquellas casillas en que los sujetos han omitido respuesta o que por

factores de diversa índole no figuran en la planilla sobre la que trabajamos. En general, dos son las opciones más seguras: excluir los casos (lo que obviamente dependerá del número total de casos que hemos incorporado al análisis, ya que nos obligará a trabajar solo con aquellos sujetos que tienen datos para todos los ítems) o reemplazar el valor perdido por la media del grupo para ese ítem (de modo que neutraliza el caso en el ítem perdido, pero sus valores para el resto permanecen utilizables). La otra posibilidad (excluir casos según pareja, utiliza todos los sujetos que tienen valores válidos para un par de variables en el cálculo de la correlación de dicho par).

Figura 43. *Análisis Factorial: Opciones*



Por otra parte, en este menú podemos pedir una determinada visualización de los datos, ordenándolos por tamaño, de modo tal que se nos presenten las saturaciones factoriales ordenadas descendentemente para cada factor. Aún más importante, es que podemos determinar el coeficiente de correlación mínimo con que queremos que nos muestren las tablas de variables. Así, cada factor extraído solo mostrará aquellos ítems o variables con valor absoluto igual o mayor que el especificado. De este modo tendremos una presentación mucho más clara de las saturaciones de cada factor.

Al revisar los resultados del procedimiento análisis factorial en el visor de resultados, nos encontraremos con una variedad de información que incluye: las medidas de adecuación muestral seleccionadas (KMO y test de esfericidad), comunalidades, tabla

para la varianza total explicada y para el procedimiento rotado, la matriz de componentes, la matriz de componentes rotada y la matriz de transformación de los componentes. En la Figura 44 podemos observar las medidas de adecuación de nuestros datos (KMO y Prueba de esfericidad de Bartlett) que nos indican qué tan buenos son los datos que tenemos como para que el análisis resulte una buena idea. El índice KMO compara los coeficientes de correlación obtenidos con las magnitudes de los coeficientes de correlación parcial y nos señala si la correlación entre los pares de ítems puede o no explicarse a partir de otros ítems. Cuando el KMO toma un valor considerado bajo se desaconseja la aplicación de un análisis factorial. En general, se considera que índices KMO de entre 0.90 y 1 son "maravillosos", entre 0.80 y 0.90 serían "meritorios", entre 0.70 y 0.80 se considerarán "medianos", entre 0.60 y 0.70 como "mediocres", entre 0.50 y 0.60 como "bajos" y entre 0 y 0.50 como "inaceptables" (García Jiménez, Gil Flores y Rodríguez Gómez, 2000). Para nuestro caso la medida KMO toma un valor de .871 (considerada meritoria).

Figura 44. *Índices de adecuación de la muestra*

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,871
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	1143,126
	gl	190
	Sig.	,000

Paso seguido observamos los resultados obtenidos por medio de la prueba de esfericidad de Bartlett, la que nos permitirá rechazar la hipótesis nula de que la matriz correlaciones es una matriz de identidad. Una matriz de identidad es aquella que contiene unos en la diagonal principal y ceros en las casillas restantes, es decir, cada ítem tiene una correlación de perfecta

consigo mismo y nula correlación con el resto de las variables. Como se observa en la Figura 44, el valor del grado de significación es de 0,000, por lo que nuestra matriz contiene correlaciones entre las diferentes variables (lo que ya pudimos apreciar al observar las matrices de correlación).

Tomados en conjunto, ambos indicadores nos autorizarían a considerar pertinente la ejecución del análisis factorial sobre nuestros datos, por lo que el paso siguiente consiste en analizar las comunalidades. Se debe tener claro que existe una serie de otros índices de adecuación que pueden consultarse, tales como el determinante (que analiza la existencia de correlaciones elevadas entre variables), la matriz de correlaciones (para visualizar el tamaño de la correlación entre los elementos de la escala), la matriz de correlación anti-imagen y el análisis de residuales (García Jiménez, Gil Flores y Rodríguez Gómez, 2000; Pardo y Ruiz, 2002; Cárdenas, 2006).

Figura 45. *Comunalidades*

Comunalidades

	Inicial	Extracción
i1	1,000	,661
i2	1,000	,776
i3	1,000	,450
i4	1,000	,660
i5	1,000	,560
i6	1,000	,660
i7	1,000	,293
i8	1,000	,629
i9	1,000	,421
i10	1,000	,567
i11	1,000	,685
i12	1,000	,576
i13	1,000	,423
i14	1,000	,623
i15	1,000	,485
i16	1,000	,654
i17	1,000	,626
i18	1,000	,641
i19	1,000	,517
i20	1,000	,526

Método de extracción:
Análisis de Componentes principales.

Se denomina comunalidad a la proporción de varianza explicada por los componentes. Para el caso de extracción por medio de componentes principales (que es el de nuestro ejemplo), las comunalidades iniciales son siempre iguales a uno. Valores cercanos a uno, posteriores a la extracción, indican que la variable queda totalmente explicada por los componentes.

El valor de la comunalidad está comprendido entre 0 y 1. Un valor cercano a cero indica que los componentes no explican casi nada de la variabilidad de una variable, mientras que los cercanos a uno indican que una alta proporción de la variabilidad queda explicada por los componentes.

En nuestro caso, la mayor parte de las comunalidades son satisfactorias (a excepción del ítem 7), por lo que podemos afirmar que los 19 ítems restantes quedan explicados adecuadamente por los componentes, puesto que no hay valores bajos, próximos a cero (Figura 45). La fluctuación de la extracción genera un rango que toma valores comprendidos entre 0,776 y 0,293.

Figura 46. *Varianza total explicada*

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	7,419	37,094	37,094	7,419	37,094	37,094	3,142	15,709	15,709
2	1,623	8,113	45,208	1,623	8,113	45,208	2,937	14,683	30,392
3	1,247	6,233	51,441	1,247	6,233	51,441	2,892	14,460	44,853
4	1,166	5,828	57,269	1,166	5,828	57,269	2,483	12,416	57,269
5	,971	4,854	62,123						
6	,920	4,601	66,724						
7	,828	4,139	70,863						
8	,792	3,961	74,825						
9	,754	3,768	78,593						
10	,617	3,083	81,676						
11	,559	2,795	84,471						
12	,513	2,564	87,036						
13	,448	2,241	89,277						
14	,427	2,134	91,411						
15	,373	1,866	93,277						
16	,343	1,717	94,994						
17	,326	1,629	96,623						
18	,273	1,364	97,987						
19	,226	1,124	99,110						
20	,178	,890	100,000						

Método de extracción: Análisis de Componentes principales.

La extracción de factores se realiza siguiendo la regla de mantener solo aquellos que tengan autovalores iniciales por encima de uno (se asume que es posible explicar el 100% de la varianza). Recordemos que estamos trabajando para nuestro ejemplo con un método de extracción de análisis de componentes principales. En la Figura 46, podemos apreciar el resultado de dicha extracción.

La columna de autovalores expresa la cantidad de varianza total que está explicada por cada factor. Podemos observar que se han extraído cuatro factores para nuestra aplicación, logrando explicar una varianza de 57,26%. El factor uno es siempre el que más peso otorga a la varianza total, en nuestro caso casi quintuplicando al segundo factor. Se pueden extraer tantos componentes como variables tengamos para analizar, pero se recomienda pedir un número menor (lo que se realiza al momento de ejecutar el procedimiento en SPSS).

La columna "sumas de las saturaciones al cuadrado de la extracción" contiene los mismos datos que su precedente, y es de bastante utilidad para otro tipo de métodos de extracción en los que puede ayudar a determinar el número idóneo de factores a extraer (en el caso del método de componentes principales ambas columnas son iguales).

La columna "sumas de las saturaciones al cuadrado de la rotación" nos muestra los porcentajes de varianza explicados por cada factor una vez realizada la rotación, y que consideraremos los porcentajes finales. Allí podemos apreciar cómo se ha realizado una ponderación para cada variable más equitativa, permitiéndonos la extracción de factores independientes pero con una mejor repartición de los ítems en su interior. Así, si el primer factor en la solución sin rotar explicaba por sí solo el 37.09% de la varianza, posteriormente a la rotación explica el 15.70%.

La Figura 47 (matriz de extracción de componentes) recoge los pesos factoriales de cada variable en los diez componentes extraídos. El peso o carga factorial nos indica el grado de correlación entre las variables y los componentes. Hemos pedido solo aquellos pesos superiores a 0,40. Por otra parte, podemos observar en esta tabla la agrupación de variables en cada factor, es decir, los reactivos que se encontrarían en cada componente. Son estos reactivos lo que le asignan sentido al factor, en tanto ellos determinan lo que agrupa este en consideración de lo que hay de común entre ellos.

Figura 47. *Matriz de extracción de componentes*

Matriz de componentes [a]

	Componente			
	1	2	3	4
i1	,640		-,484	
i2	,754		-,411	
i3	,512			
i4	,713			
i5	,648			
i6	,561		,560	
i7		,436		
i8	,609			
i9	,593			
i10	,632	,428		
i11	,708			
i12	,652			
i13	,578			
i14	,612			,407
i15	,600			
i16	,777			
i17	,460			,579
i18	,511	,454	,414	
i19	,571			-,433
i20	,564			

Método de extracción: Análisis de componentes principales.

a. 4 componentes extraídos

El paso siguiente ha sido realizar una rotación de los componentes para hacer más sencilla su interpretación (se comprenderá que la rotación no es necesaria para aquellos casos en que la interpretación del análisis sea sencilla). Hemos elegido una rotación ortogonal VARIMAX, que permite rotar los ejes de referencia del origen de forma de lograr redistribuir la varianza logrando un patrón de factores más significativo, y con la intención de simplificar al máximo los vectores columna

de la matriz. Es sobre esta matriz resultante (Figura 48) desde donde hemos decidido seleccionar los ítems correspondientes a cada factor.

Figura 48. *Matriz de componentes rotada*

Matriz de componentes rotados [a]

	Componente			
	1	2	3	4
i1			,753	
i2			,775	
i3		,402		,482
i4		,656		
i5				,609
i6		,720		
i7	,533			
i8	,630		,468	
i9			,521	
i10	,694			
i11	,535			,517
i12	,632			
i13		,412	,411	
i14				,677
i15		,545		
i16	,589			
i17				,699
i18	,682			
i19		,521	,407	
i20		,658		

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 8 iteraciones.

Finalmente, SPSS nos entrega la matriz de transformación de los componentes, la que nos indica el grado de correlación existente entre los diferentes factores extraídos. Esta es la matriz que se ha utilizado para rotar la solución inicial. A esta última figura no deberemos dedicarle mucho interés, en consideración de nuestros fines.

Figura 49. *Matriz de transformación*

Matriz de transformación de las componentes

Componente	1	2	3	4
1	,512	,518	,513	,454
2	,855	-,384	-,249	-,245
3	,072	,567	-,801	,178
4	-,043	-,513	-,181	,838

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

8.8.2. Análisis de Conglomerados (cluster)

Es un procedimiento de reducción de datos que tiene como objeto clasificar grupos de objetos basándose en una serie de atributos o características. De modo más específico, el análisis de conglomerados (o clusters) trabaja sobre un conjunto de datos que tienen información sobre una serie de objetos o unidades de análisis e intenta organizarlos en una serie reducida de grupos formados por objetos relativamente homogéneos. Para ello se clasifican las unidades de análisis de modo que cada unidad sea lo más similar posible respecto de las que están en el mismo conglomerado en relación a algún criterio (Martínez Arias, 1999). La idea es que cada conglomerado resulte ser muy homogéneo internamente y con alta heterogeneidad hacia el exterior (respecto de los otros conglomerados). De este modo, por medio del AC, se logra desarrollar subgrupos significativos de individuos u objetos, los que servirán para clasificar una muestra de entidades reduciéndola a un número pequeño de grupos excluyente basados en las similitudes de dichas entidades (Hair, Anderson, Tatham y Black, 2000).

El análisis de conglomerados (AC) se realiza sobre la similaridad entre objetos. Para ello ha de construirse desde los datos una matriz de similaridad que funciona como medida de correspondencia o parecido entre objetos. Esta medida de similitud entre objetos puede evaluarse a través de medidas correlacionales, medidas de distancia o de asociación (Martínez Arias, 1999), y su objetivo es definir la estructura de los datos colocando las observaciones más parecidas en grupos.

El AC es útil para cuantificar las características estructurales de un conjunto de observaciones. Esta cuantificación es calculada desde la matriz de similitud y se objetiva en la formación de grupos o partición.

Existen dos grandes grupos de procedimientos aglomerativos: los métodos jerárquicos y los iterativos. El primer grupo es el que revisaremos en este apartado y supone la construcción de una jerarquía de estímulos en forma de árbol (dendograma), en la cual los resultados de un estadio temprano están siempre anidados dentro de otro posterior. Estos constan a su vez de dos tipos básicos, los aglomerativos (cada objeto es inicialmente considerado como un conglomerado separado, y en los pasos siguientes, los objetos más próximos son combinados en nuevos conglomerados, reduciendo su número a cada paso del análisis hasta llegar a un conglomerado único) y los divisivos (en este tipo el proceso es el inverso, se comienza con un gran conglomerado, que contiene todas las observaciones, que se va subdividiendo en los pasos posteriores). En los procedimientos iterativos no se construyen estructuras de árbol, sino que se asignan los objetos una vez que se determina el número de grupos (Martínez Arias, 1999).

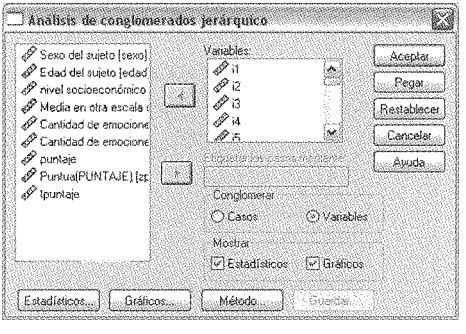
Estos dos tipos de procedimientos de conglomerados para la agrupación de los datos han sido también denominados procedimiento de K medias y análisis de conglomerados jerárquicos. Este último es el idóneo para determinar el número óptimo de conglomerados existente en los datos y el contenido de los mismos, pudiendo tanto casos como variables y elegir entre una amplia variedad de métodos de aglomeración y de medidas de distancia. Los elementos más próximos se agrupan en conglomerados a partir de su distancia. El conglomerado resultante resulta indivisible a partir de su formación (de ahí el nombre de jerárquicos). El AC de K medias permite procesar un número ilimitado de casos, pero solo permite utilizar un método de aglomeración y requiere que se proponga previamente el número de conglomerados que se desea obtener (Pardo y Ruiz, 2002).

La interpretación de los datos es sencilla, pues se trata de examinar los componentes de cada conglomerado y se le asigna

un nombre a partir de sus componentes y guiado por consideraciones teóricas, esto es, se describen las personas o variables para determinar su composición.

Para realizar dicho procedimiento debemos seleccionar la opción conglomerados jerárquicos de la opción clasificar contenida en el menú Analizar. De este modo, accedemos al cuadro de diálogo Análisis de conglomerados jerárquico (Figura 50), el cual nos muestra la lista completa de las variables, las que debemos trasladar al cuadro de "variables".

Figura 50. *Análisis de Conglomerados jerárquico*



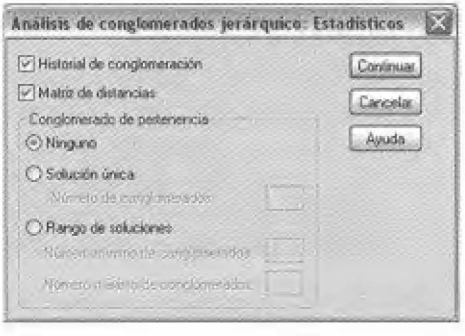
Este cuadro de diálogo nos permite decidir sobre el tipo de resultados que queremos que nos muestre el visor (ya sea agrupando casos o variables), así como permitiendo mostrar las opciones de "Estadísticos" y "Gráficos" que vienen seleccionadas por defecto. La primera nos entregará los resultados numéricos y la segunda los gráficos (historial de conglomeración y diagrama de témpanos). Es recomendable siempre trabajar con ambas opciones activadas.

Al pulsar el botón Estadísticos (Figura 51) nos aparece una serie de posibilidades como la de activar o desactivar el Historial de conglomeración (que es la tabla que nos informa sobre los elementos que se van fundiendo en cada etapa y sobre la

distancia existente entre estos al momento de la fusión, así como sobre las etapas previas y posteriores) y la Matriz de distancias (ya habíamos comentado más arriba que este procedimiento se realizaba sobre la base de la cercanía o distancia entre elementos).

El apartado del cuadro referido al Conglomerado de pertenencia permite presenciar la tabla de igual nombre y que informa sobre el conglomerado al que han sido asignados cada uno de los casos en cada etapa del análisis. Aquí, en general, mantendremos la opción que aparece por defecto y que es Ninguno.

Figura 51. *Análisis de Conglomerados jerárquico: Estadísticos*



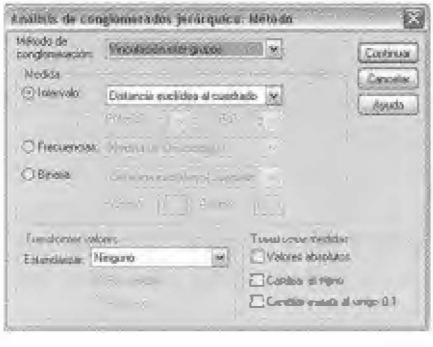
Por otra parte, al pulsar el botón "Gráficos" (Figura 52) accedemos al cuadro que nos permite obtener el dendograma, que es un gráfico que combina el diagrama de témpanos y el historial de conglomeración, mostrando mediante trazos horizontales los conglomerados y las etapas de fusión en los verticales. Este procedimiento se realiza en una escala estandarizada de 25 puntos, de modo que se transforman las distancias originales de modo proporcional y se adaptan a esta nueva medida. El dendograma es un gráfico fundamental para nuestros fines, ya que nos facilitará enormemente la interpretación de los resultados.

Figura 52. *Análisis de Conglomerados jerárquico: Gráficos*

El apartado denominado Témpanos permite manejar algunas de las características de dicho gráfico, a saber: su orientación, la representación de un conjunto de rango de soluciones o la no presentación de dicho gráfico. Aquí también mantendremos la orientación por defecto, pero podemos modificar la orientación para hacer la observación de este gráfico más amable.

Finalmente, al pulsar el botón "Método" (Figura 53), podemos acceder a la selección del método de conglomeración y al tipo de medida que utilizaremos para evaluar la distancia entre los elementos.

Figura 53. *Análisis de Conglomerados jerárquico: Método*



El método de conglomeración es un procedimiento mediante el cual es posible recalcular las distancias entre los nuevos elementos en cada etapa del proceso de fusión (Pardo y Ruiz, 2002). Los principales métodos de conglomeración son: *Método de vinculación por el vecino más próximo* (se funden los elementos de la matriz de distancias que se encuentran más próximos y la distancia entre este conglomerado respecto de los restantes elementos se calcula como la menor de las distancias entre cada elemento del conglomerado y el resto de los elementos de la matriz), *Método de vinculación por el vecino más lejano* (la distancia entre los dos conglomerados se calcula como la distancia entre sus dos elementos más lejanos), *Método de vinculación intergrupos* (aprovecha la información de todos los miembros de los dos conglomerados que se comparan, de modo que utiliza la distancia promedio existente entre todos los pares de elementos de ambos conglomerados), *Método de Ward* (permite que la pérdida de información resultante de la fusión sea mínima al trabajar sobre un cálculo de distancias mediante la suma de cuadrados), *Método de agrupación de centroides* (la matriz de distancias original solo es utilizada en la primera etapa, luego esta se va actualizando según la distancia entre sus vectores de medias) y *Método de agrupación de medianas* (permite que a la hora de caracterizar los conglomerados los más pequeños tengan la misma importancia que los grandes). Para una más completa definición y descripción existe abundante material disponible (Martínez Arias, 1999; Pardo y Ruiz, 2002; Cárdenas, 2006).

Respecto de las medidas de distancia que utilizaremos, debemos tener siempre presente que estas dependen del tipo de datos que estamos utilizando para nuestras indagaciones. En general las más utilizadas son las medidas de intervalo (para datos escalados) y las de tipo dicotómico (presencia o ausencia de atributos). Este último tipo de medida supone que los elementos son tanto más similares entre sí tanto mayor sea el número de presencia o ausencias que comparten (aunque

no necesariamente las co-presencias o co-ausencias tienen el mismo valor informativo). Algunas de las medidas de distancia más comunes son: *Distancia euclídea, Distancia euclídea al cuadrado, Diferencia de tamaño* (estas tres primeras son medidas de disimilaridad cuyo valor mínimo es 0, pero que no tienen máximo), *Diferencia de configuración* (medida de disimilaridad que toma valores entre 0 y 1), *Varianza* (medida de disimilaridad con valor mínimo 0, pero sin máximo), *Dispersión* (medida de similaridad que toma valores entre 0 y 1), Forma (medida de disimilaridad sin límites inferior ni superior), *Coeficiente de Phi de cuatro puntos* (es la medida de similaridad más utilizada para datos binarios y considerada la versión binaria del coeficiente de correlación de Pearson que toma valores entre –1 y 1), *Lambda de Goodman y Kruskal* (medida de similaridad que evalúa el grado en que el estado presente o ausente de una característica en un variable puede predecirse a partir del estado de otra. Toma valores entre 0 y 1), *Jaccard* (medida de similaridad que no tiene en cuenta las ausencias conjuntas y pondera por igual las concordancias de las discordancias) y, finalmente, la *Q de Yule* (medida de similaridad que toma valores entre –1 y 1).

Una vez realizado el procedimiento completo para el análisis de conglomerados pulsamos el botón aceptar del cuadro de diálogo principal y ya podremos dirigirnos al visor de resultados de SPSS para observar e interpretar los resultados.

Respecto del análisis de conglomerados, hemos realizado un procedimiento de conglomerados jerárquicos, el que nos ha permitido aglomerar las variables (escogiendo tanto el método de aglomeración y la medida de distancias) y proceder de modo jerárquico. Respecto del método de aglomeración elegido, podemos decir que hemos utilizado el método de vinculación intergrupos o promedio, debido a que con él se puede aprovechar la información de todos los miembros de los conglomerados que se comparan (Pardo y Ruiz, 2002). La

distancia entre conglomerados se calcula como la distancia promedio existente entre todos los pares de elementos de cada conglomerado. En la Figura 54 se puede observar el historial de conglomeración, esto es, todo el proceso de conglomeración etapa por etapa, que en nuestro caso son 19. En cada etapa se unen o funden dos elementos.

Figura 54. *Historial de conglomeración*

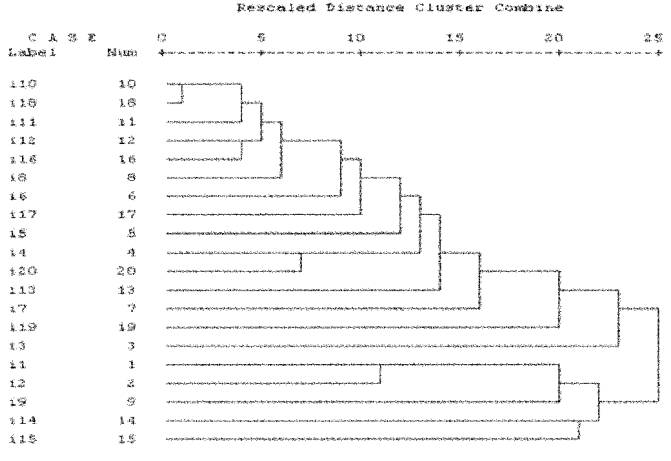
Historial de conglomeración

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	10	18	111.000	0	0	3
2	12	16	163.000	0	0	4
3	10	11	165.500	1	0	4
4	10	12	182.167	3	2	5
5	8	10	201.200	0	4	7
6	4	20	211.000	0	0	11
7	8	9	249.667	0	5	8
8	5	17	253.143	7	0	10
9	1	2	280.000	0	0	15
10	5	6	290.250	0	8	11
11	4	5	296.611	6	10	12
12	4	13	313.455	11	0	13
13	4	7	343.500	12	0	14
14	4	19	410.923	13	0	18
15	1	9	416.000	9	0	17
16	14	15	421.000	0	0	17
17	1	14	450.500	15	16	19
18	3	4	453.571	0	14	19
19	1	3	497.533	17	18	0

La columna de coeficientes nos informa sobre el valor que asume la distancia en la que se encuentran los casos antes de la fusión (dato que también puede observarse en la matriz de distancias). La columna "etapa en que el conglomerado aparece por primera vez" recoge la etapa en que se están formando los conglomerados que se están fundiendo en cada momento (un valor de 0 indica que es un caso individual y su primera aparición para fusión, un valor superior indica la etapa en que se formó el conglomerado). Finalmente, la columna "próxima etapa" nos informa sobre la etapa en que el conglomerado recién formado se volverá a fusionar.

Todo este proceso de fusión que hemos observado en la figura anterior puede observarse en el dendograma (figura 55), que es un gráfico en el que quedan representadas las etapas del proceso de fusión y las distancias existentes entre los elementos fundidos en cada etapa. Eso sí, estas distancias han sido reescaladas sobre una medida estandarizada de 25 puntos, lo que en nuestro caso significa que el conglomerado final que tiene una distancia de 25 le corresponde una distancia inicial de 497.53 y al conglomerado inicial con valor asignado de 1 le corresponde una distancia de 111.0 (tal como puede observarse en el historial de conglomeración).

Figura 55. *Dendograma*



Como se aprecia en el gráfico, las líneas verticales identifican los elementos fundidos o conglomerados, y las horizontales indican la distancia entre ellos (que como ya se indicó se realiza sobre una escala de 25 puntos). De este modo, las fusiones realizadas cerca del origen de la escala (hacia el lado izquierdo del gráfico) indican que el conglomerado es bastante homogéneo. A la inversa, mientras más alejado del origen (hacia la derecha del gráfico) mayor heterogeneidad. Podemos de este

modo apreciar qué ítems de la escala que hemos venido utilizando son considerados por la muestra utilizada como más semejantes entre sí. Así, hemos combinado la información de los gráficos anteriores, aunque alterando la escala original de distancias para hacerlo más fácilmente interpretable (pero manteniendo la proporción de las distancias de fusión representadas por las líneas horizontales), al permitirnos evaluar la homogeneidad de los conglomerados y facilitarnos la decisión sobre el número óptimo de estos. Este es el gráfico que se entrega al momento de la exposición de los resultados.

8.8.3. Escalamiento Multidimensional

El escalamiento multidimensional es un procedimiento que permite determinar las imágenes subjetivas asociadas a un conjunto de objetos por parte de los sujetos y las dimensiones sobre las que se basan esos juicios. Para conseguir esta representación, la técnica parte de los juicios de similitud o preferencia sobre objetos expresados por los sujetos. Como el AC, esta técnica también parte desde una matriz de similitud, las cuales son transformadas en distancias, permitiendo situarlas en un espacio multidimensional (Martínez Arias, 1999). Suele ser útil cuando se pretende poner de relieve las dimensiones latentes que subyacen al juicio de los sujetos y para obtener evaluaciones comparativas de objetos entre los cuales no existe un criterio definido de comparación. De este modo, el MDS permite, a través de los mapas perceptuales, representar las proximidades relativas entre un conjunto de objetos o estímulos como distancias en un espacio de baja dimensionalidad (Hair, Anderson, Tatham y Black, 2000; Real Deus, 2001).

El objetivo del MDS es determinar el número de dimensiones que explican los juicios de los sujetos y las coordenadas o puntos de dichas dimensiones. Se debe manejar el supuesto de que los sujetos no conceden la misma importancia a todas las

dimensiones incluso aunque todos las perciban. Por otra parte, se debe tener claro que los juicios de los sujetos no tienen que permanecer estables es el tiempo. En términos más generales, con el MDS se logran cuatro objetivos generales: determinar las dimensiones que utilizan los encuestados cuando evalúan los objetos, determinar el número de dimensiones que pueden utilizarse en una situación particular, determinar la importancia relativa de cada dimensión y definir como se relacionan perceptualmente los objetos (Hair, Anderson, Tatham y Black, 2000). Todo ello permite transformar, como ya se ha dicho, los juicios de similitud o preferencia en distancias representadas en un espacio multidimensional. Todo esto se realiza basado en el supuesto de que la comparación de objetos se hace sobre dimensiones (ya sean estas objetivas o subjetivas).

La distancia observada, resultante de la aplicación del procedimiento será baja entre los elementos más parecidos y alta entre aquellos menos similares. En todos los casos las proximidades (que son nuestros datos de entrada) son las que indicarán la cercanía o lejanía, ya sea objetiva o subjetiva, entre los objetos. Cada dimensión queda graficada por un eje que nos indica la escala que el sujeto o grupo ha utilizado para estimar las proximidades, así como el posicionamiento de los estímulos en cada una de estas (Real Deus, 2001). El MDS nos permite, de este modo, interpretar la estructura latente en forma de dimensiones o de agrupaciones significativas de estímulos. Cuando nuestros datos de entrada no son proximidades, deberemos obtener a partir de ellos un coeficiente de similaridad. Es decir, si con el análisis de conglomerados sabíamos algo acerca de los juicios de similitud realizados por los sujetos de una muestra, con el escalamiento logramos algunas pistas para entender la base sobre la cual se realizan dichos juicios de similitud.

Los criterios de ajuste utilizados en este procedimiento son el stress y RSQ (correlación múltiple al cuadrado), siendo el

primero un indicador de mal ajuste (debiendo ser por ello lo más bajo posible, idealmente bajo 0.10) y el segundo, de bondad de ajuste (mejor cuanto más cercano a uno sea su valor) (Martínez Arias, 1999; Real Deus, 2001). El ajuste de los datos también puede graficarse por medio de un diagrama de dispersión. En todo caso, el valor del stress dependerá del número de dimensiones (tiende a bajar a mayor dimensionalidad) y de estímulos (a más estímulos mayor stress). El tipo de MSD empleado dependerá de una serie de factores tales como el tipo de datos de entrada o el número de matrices de proximidad empleados. A partir de estos elementos se utilizará un modelo métrico (datos medidos en escala de intervalo) o no métrico (datos medidos en escala ordinal), el que a su vez puede tener una o varias matrices de entrada, tratadas como replicaciones o como representaciones de un mismo espacio ponderadas de forma diferente (MDS replicado, modelo INDSCAL y modelo GEMSCAL). Otros modelos de MSD realizan operaciones sobre matrices de proximidad asimétricas, esto es, en las que la distancia entre el objeto a y b no es igual a la que hay entre el objeto b y el a (modelo ASCAL) o entre varias matrices de proximidad asimétricas (modelo AINDS). Sobre los rasgos distintivos y aplicaciones de cada uno de estos modelos hay abundante documentación disponible (Real Deus, 2001).

Para la interpretación de los resultados se suelen examinar las dimensiones utilizadas para ubicar los objetos en el mapa y sus respectivos valores numéricos en las coordenadas de los estímulos. También pueden realizarse análisis de regresión múltiple para las distintas dimensiones de valoración. Pero aún más básicamente, consiste en describir, interpretando, las dimensiones preceptúales y su correspondencia de atributos. Para realizar el procedimiento de MDS por medio de SPSS 14.0 se comienza por la selección de la opción Escalamiento multidimensional de la opción Escalas de menú Analizar (Figura 56). De este modo se desplegará en la pantalla el cuadro de

diálogo de Escalamiento multidimensional, el que nos muestra en el lado izquierdo la lista completa de variables de tipo numérico, y que como en el caso del procedimiento anterior trasladamos hacia la sección rotulada como variables.

Figura 56. *Escalamiento multidimensional*

Como los datos que estamos utilizando son distancias no debemos señalar la opción *Crear distancias a partir de los datos*, ubicada en la sección de igual nombre. En caso de que nuestros datos fueran de tipo binario deberíamos especificarlo y solo entonces aparecería disponible el botón etiquetado como "Medida" y que nos permite crear una medida de proximidad a partir de nuestros datos de perfil. Los tipos de medida de distancia disponibles en el programa son: Distancia euclídea (es la opción por defecto del programa y la que nosotros utilizaremos para nuestros datos), distancia euclídea al cuadrado, diferencia de tamaño, diferencia de configuración, varianza y Lance y Williams. Además, debemos especificar los valores de la dicotomía (en nuestro caso son 0 para indicar ausencia y uno para la presencia del atributo o palabra).

Al pulsar las opciones referidas al "Modelo" (Figura 57) nos encontramos con la especificación del nivel de medida de nuestros datos (ordinal, intervalo o razón), con la opción para

fijar el número de dimensiones que se desea obtener en la gráfica y de selección del modelo de escalamiento. El modelo de escalamiento variará según sea el tipo de datos de entrada que utilicemos. Para nuestro caso el modelo puede considerarse métrico, es decir, supone una relación lineal entre proximidades y distancias. En el caso de los modelos no-métricos no existen desventajas respecto del métrico, a condición de que el número de estímulos sea suficientemente elevado (Real Deus, 2001).

Figura 57. *Escalamiento multidimensional: Modelo*

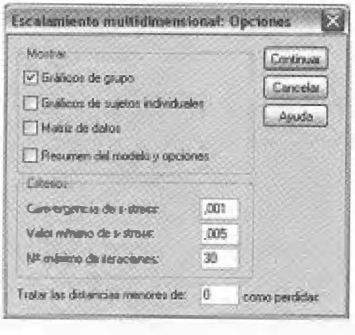
El procedimiento de escalamiento nos ofrece una serie de "Opciones" (Figura 58) de visualización. La que nos interesa sobre todas es la referida al "Gráfico de grupos" que nos aporta el mapa realizado sobre los estímulos sobre los ejes o dimensiones que hemos especificado (y que en nuestro caso son dos). Además, podremos pedir la incorporación de alguna información adicional y alterar los criterios de convergencia del análisis. Es el momento en que podremos especificar el tipo de gráfico que deseamos, ya sea este grupal o para los sujetos individuales. El apartado referido a los *Criterios* se mantendrá inalterado, y solo se retendrá por ahora la idea de que estos criterios dicen referencia con el número de iteraciones máximas dentro de las cuales se pretende obtener la minimización de uno de los índices de ajuste, el denominado s-stress.

Una vez que se ha pedido mostrar el gráfico de grupo, retornamos al cuadro de diálogo principal, pulsamos el botón

aceptar y nos movemos hacia el visor de resultados para observar e interpretar los resultados obtenidos por medio de este procedimiento.

El escalamiento multidimensional (MDS) nos permitirá llevar las distancias obtenidas a un espacio de baja dimensionalidad, de modo que representaremos las proximidades entre objetos como distancias entre puntos en un espacio bidimensional (mapa). Pero además, al realizar este procedimiento, obtenemos los ejes o dimensiones que los sujetos de la muestra han utilizado para estimar las proximidades, permitiéndonos una interpretación de los factores que pueden subyacer a las relaciones establecidas. A mayor cercanía estimada entre estímulos, por el grupo de sujetos, encontraremos una mayor valoración de parecido. Esto es, si la similitud entre estímulos es juzgada como alta, entonces las distancias en el mapa bidimensional serán bajas. Se hace posible de este modo observar la estructura oculta de los datos.

Figura 58. *Escalamiento multidimensional: Opciones*



La Figura 59 nos entrega las coordenadas obtenidas para los veinte estímulos utilizados. Se observa que los índices de ajuste Stress y RSQ toman los valores .192 y .852 respectivamente. Como se sabe, el primero es un índice de "mal ajuste" y por lo tanto se esperan valores lo más cercanos posibles a 0, mientras que el segundo (RSQ), es indicador de bondad de ajuste y por ello se desea que sea lo más cercano posible a

1. En nuestro caso, tanto los índices de Stress y RSQ nos indicarían que habría algunos leves problemas con el ajuste de los datos (esta leve alza podría ser debida al elevado número de estímulos utilizados). Tomados en su conjunto los indicadores nos permiten continuar con nuestro análisis.

Figura 59. *Coordenadas para los estímulos*
El ajuste de nuestros datos a las distancias derivadas a partir

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	2
1	i1	2,9773	,4077
2	i2	,8512	,2847
3	i3	,2794	-1,7679
4	i4	-,3512	-,4251
5	i5	-,2249	-,3749
6	i6	-1,3175	-,1950
7	i7	-,9192	,9193
8	i8	-,2152	,7045
9	i9	1,6474	1,4841
10	i10	-1,3060	,2484
11	i11	-,1545	-,0257
12	i12	-,5828	,4656
13	i13	-1,4028	-,1000
14	i14	1,5568	-,9529
15	i15	1,6757	-1,2614
16	i16	-,1559	-,0291
17	i17	-1,3391	-,4935
18	i18	-1,4788	-,0437
19	i19	,6708	,7661
20	i20	-,2097	,3877

For matrix
Stress = ,19205 RSQ = ,85250

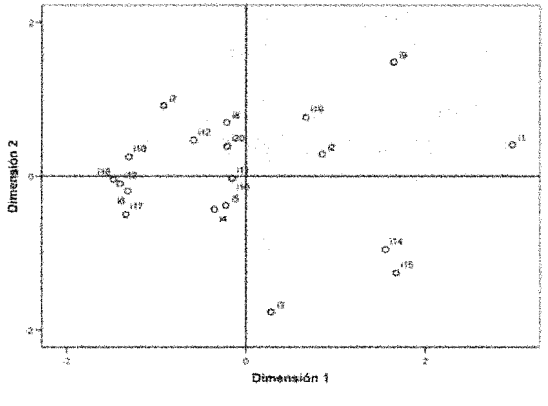
de la matriz de coordenadas también puede ser apreciado visualmente por medio del diagrama de dispersión que relaciona las proximidades contenidas en la matriz de datos con las distancias existentes entre las variables representadas. En este gráfico de dispersión no se muestran los valores originales de las disimilaridades, sino la transformación lineal de estas (disparidades). El ajuste de los datos suele ser mejor en aquellos casos en que la distancia entre estímulos es mayor debido a que el índice S-stress busca el mejor ajuste entre disparidades y distancias al cuadrado.

En la figura 60 se pueden observar los resultados obtenidos por medio del procedimiento escalamiento multidimensional. La interpretación de dicho gráfico debe realizarse tomando cada

una de las dimensiones que hemos pedido por separado (que volvemos a recordar, en nuestro caso de dos) y que corresponden a los ejes vertical y horizontal del gráfico. De modo que todas los ítems utilizados deben ubicarse en el continuo que dicho eje representa, de modo tal que se pueda inferir a partir de dichos ítems que dicho eje opone cual es la categoría subyacente que se está utilizando para realizar los juicios del grupo. No hay que olvidar que dentro del gráfico la cercanía entre estímulos es considerada como una atribución de similitud, por lo que se procede de modo de designar con un nombre el eje completo, teniendo en cuenta los estímulos que deja agrupados en cada polo.

La interpretación de la dimensión uno del gráfico implicaría apreciar la opción entre los ítems de la izquierda (I10, I13, I16, I17 e I18) versus los de la derecha (I1, I9, I14 e I15). Este eje opone dos conjuntos de ítems que nos deberían de informar sobre las similitudes percibidas hacia el interior de los mismos y que nos entregarían una clave para comprender el criterio utilizado para realizar dicha distinción. Lo mismo ocurriría con el eje dos, que opondría los ítems ubicados en la parte superior del mapa (I9, I7, I8 e I19) respecto de aquellos ubicados en la parte inferior (I3, I14 e I15).

Figura 60. *Escalamiento Multidimensional: Gráfico de las dimensiones*



De este modo, hemos conseguido que los sujetos clasifiquen una serie de estímulos sobre los que nos interesaba tener su opinión, en términos de distancia percibida. Esto ya nos ha entregado una valiosa información, pero además – y quizás más importante aún- nos ha indicado cuáles son los criterios utilizados para realizar dicha clasificación.

8.9. Bibliografía

- Cárdenas, M. (2006). El análisis multivariante de las representaciones sociales. Antofagasta: Editorial Universidad Católica del Norte.

- García Jimenez, E., Gil Flores, J., y Rodríguez Gómez, G. (2000): Análisis factorial. Cuadernos de estadística 7. Madrid: Editorial La Muralla / Editorial Hespérides.

- Hair, J., Anderson, R., Tatham, R. Y Black, W. (2000): Análisis multivariante. Madrid: Prentice Hall, Pearson Educación.

- Martínez Arias, R. (1999): El análisis multivariante en la investigación científica. Cuadernos de estadística 1. Madrid: Editorial La Muralla / Editorial Hespérides.

- Pardo, A. & San Martín, R. (2004). Análisis de datos en psicología II. Madrid: Pirámide.

- Pardo, A. y Ruiz, M. A. (2002). SPSS 11. Guía para el análisis de datos. Madrid: McGraw-Hill.

- Real Deus, J. E. (2001): Escalamiento multidimensional. Cuadernos de estadística 14. Madrid: Editorial La Muralla / Editorial Hespérides.

- Tejedor, F. J. (1999). Análisis de Varianza. Cuadernos de Estadística 3. Madrid: Editorial la Muralla.