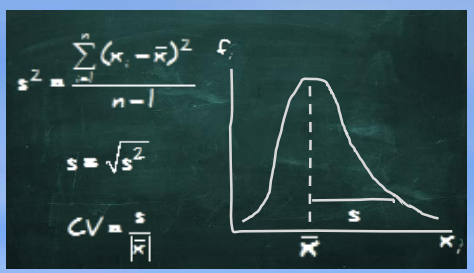
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.2
Preparación de los datos para el análisis

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2015). Preparación de los datos para el análisis. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. Capítulo III.2. 1ª edición. Edición digital: http://ddd.uab.cat/record/129381

Capítulo redactado en febrero de 2015

Índice de contenidos

PARTE III
Capítulo 2

Preparación de los datos para el análisis

Los datos que se manejan en la investigación social habitualmente requieren que sean preparados para su análisis. Esta necesidad se puede dar desde el inicio o durante el proceso mismo de análisis e interpretación de la información. Cuando nos referimos a la preparación de los datos entendemos que se trata de un conjunto de tareas de procesamiento de los datos que engloba desde el registro y la identificación en un soporte informático, pasando por la depuración de los mismos, y su transformación, que incluye tanto la modificación de la información original como la creación do otra nueva a partir de las variables existentes, o el tratamiento de ficheros de datos.

Preparar los datos para el análisis seguramente es una de las tareas menos reconocidas y a la vez de las más importantes en la investigación. Quizás porque suele ser una tarea más técnica que se suele dejar en manos de hábiles especialistas en el manejo de los programas informáticos. Pero la calidad de los datos depende enormemente de este conjunto de aspectos en interrelación con las demás fases del proceso de investigación.

La matriz de datos original que se obtiene en un proceso de investigación es pues un material informativo bruto que requiere su adaptación y acondicionamiento a las necesidades de explotación y análisis de los datos. Estas operaciones se realizan con la ayuda del software específico de tratamiento y análisis de los datos con el que se trabaje. En el Gráfico III.2.1 se presenta el organigrama del proceso de datos que resume y esquematiza la dinámica de trabajo general con el software para realizar las distintas tareas de preparación de los datos para el análisis. Se presenta haciendo referencia en particular a matrices de datos y programas de sintaxis en SPSS, pero es aplicable como dinámica igualmente al trabajo con R o SPAD.

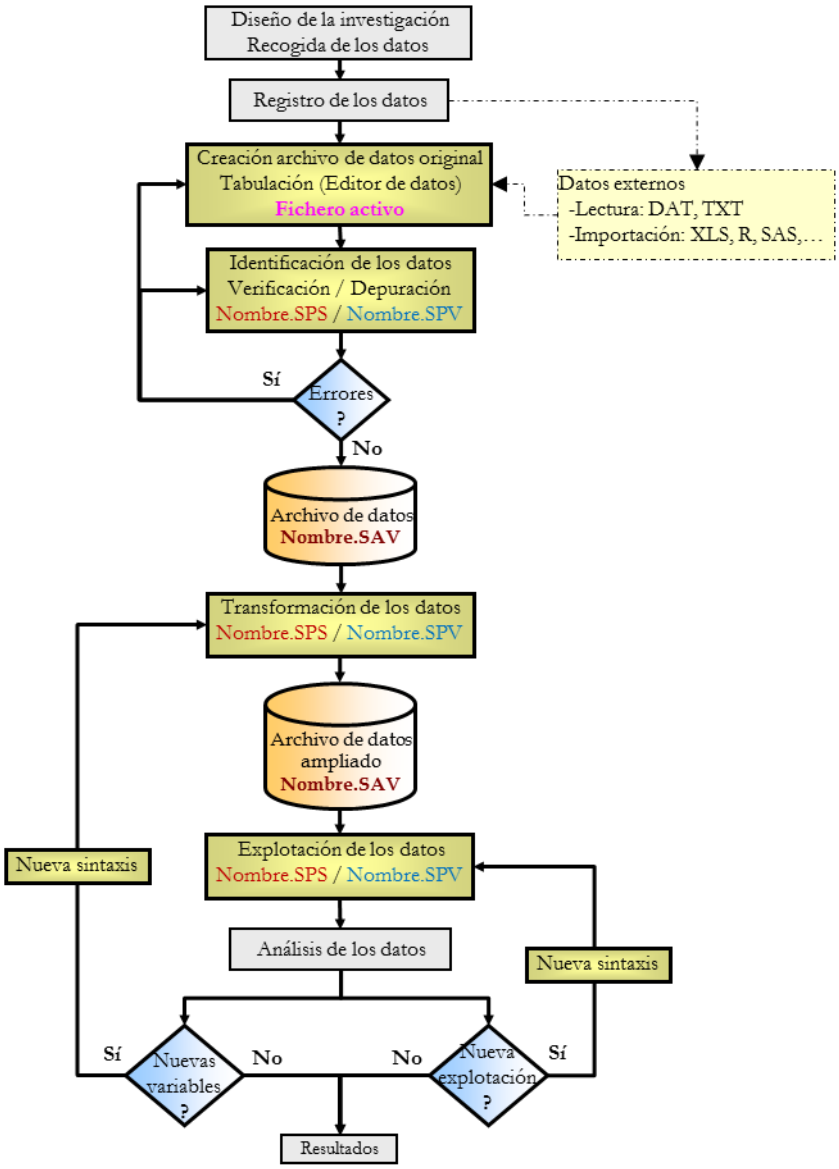
El proceso de datos implica cuatro tareas fundamentales:

1) Creación e identificación los datos, ya sea grabándolos ("picándolos") nosotros mismos[1], o importándolos por medio de la lectura de archivos de datos externos de formato plano (TXT, DAT) o con formatos de otros sistemas (XLS, SAS, R,…).

[1] Existe software específico para esta tarea como el Data Entry en SPSS que permite elaborar plantillas de introducción, identificación y control de la grabación de los datos.

Se genera así el fichero activo del sistema que grabaremos en el disco duro con un nombre identificativo.

Gráfico III.2.1 Organigrama del proceso de datos con SPSS



2) Verificar la corrección de los datos y de su identificación para depurarlos (corregirlos) en el caso de que detectáramos errores[2]. Distintos comandos se pueden utilizar para detectar primero y corregir después los datos erróneos.
3) Transformar los datos originales con el objetivo de acondicionar las variables para su explotación y análisis, tarea que conlleva habitualmente la generación de nuevas variables que amplían la matriz de datos original[3]. La transformación de los datos

[2] Buena parte de la depuración puede o debe realizarse también en la fase previa de trabajo campo, como en el caso de una encuesta. Los sistemas de recogida asistidos por ordenador reducen enormemente este trabajo.
[3] En algunos procesos de investigación por encuesta las matrices originales de datos con un número dado de variables pueden verse duplicadas al final del proceso.

puede entenderse también en relación a las tareas de manipulación de la matriz de datos como conjunto (ponderando, seleccionando, ordenando, agregando,… los datos) o fusionando un fichero de datos con otros.

4) El análisis de los datos a partir de su explotación con los diferentes procedimientos de tabulación y análisis estadísticos (univariables, bivariables y multivariables) orientados por los objetivos de la investigación y el modelo de análisis.

En este capítulo nos dedicaremos a dar cuenta de las tres primeras tareas. Con ellas alcanzaremos a conocer la calidad, estructura y propiedades de los datos que manejamos. A partir del siguiente capítulo veremos los distintos procedimientos de análisis, teniendo en cuenta que habitualmente conllevan también la necesidad de realizar nuevas transformaciones de los datos como se ilustra en el organigrama. Veremos esas tareas con SPSS y R, después de una presentación de sus características, y las ejemplificaremos con distintos ejercicios prácticos de tratamientos de datos.

1. Creación e identificación de los datos

Como hemos comentado los datos se pueden crear a partir dos procedimientos básicos: los grabamos o los importamos. Los datos así creados constituyen la matriz de datos, un conjunto de filas y columnas que responden a unos criterios de codificación de la información. Estos criterios y otros aspectos que los caracterizan nos permiten identificarlos y generar lo que denominamos como el diccionario de los datos.[4]

Realizaremos un ejercicio práctico de creación de una sencilla matriz de datos introduciendo los datos y después otros ejercicios que implican la importación de datos existentes de otras aplicaciones o formatos.

Para el primer ejercicio se considerará la información que se obtiene de las respuestas a las preguntas del cuestionario de encuesta que se adjunta en el Cuadro III.2.1. En el ejercicio se implica el proceso de codificación, de grabación y de identificación de los datos. En los apartados siguientes se detallará cómo realizar las tareas de grabación e identificación con SPSS y R. En lo que sigue presentaremos el cuestionario y un ejercicio de codificación de los datos de un caso concreto.

El cuestionario adjunto da lugar a 16 variables, cada una de las informaciones que se derivan de cada pregunta, más una primera variable adicional que identifica el número de cuestionario asignado a cada persona que responde. Llamamos a estas variables, por ejemplo: ID, P1, P2, P3_1, P3_2, P3_3, P4, P5, P6_1, P6_2, P6_3, P6_4, P6_5, P6_6, P6_7 y P7.

La matriz de datos tendrá, por tanto, 16 columnas con las respuestas de cada individuo. Estas respuestas se codifican con valores numéricos o textuales según el tipo de variable.

[4] Para ampliar la información se pueden consultar los capítulos 3, 4, 5 y 6 del manual del sistema central (IBM Corporation, 2015).

Cuadro III.2.1. Cuestionario para el ejercicio de creación de una matriz de datos

Número de cuestionario __ __ __

1. ¿Cuántos años tiene? __ __ __ No contesta ☐ (999)

2. ¿Cuál es su sexo? Varón ☐ (1)
Mujer ☐ (2)

3. ¿Me puede decir el nivel de estudios más alto que ha cursado y acabado, así como el de sus padres?

	Ego	**Padre**	**Madre**
Sin estudios, primarios inacabados	☐ (1)	☐ (1)	☐ (1)
EGB, bachillerato elemental, ESO	☐ (2)	☐ (2)	☐ (2)
Bachillerato superior, BUP, COU	☐ (3)	☐ (3)	☐ (3)
Formación Profesional			
De primer grado, oficialías	☐ (4)	☐ (4)	☐ (4)
De segundo grado, maestría industrial	☐ (5)	☐ (5)	☐ (5)
Universitarios	☐ (6)	☐ (6)	☐ (6)
No sabe	☐ (8)	☐ (8)	☐ (8)
No contesta	☐ (9)	☐ (9)	☐ (9)

4. ¿Cuál era su situación laboral la semana pasada?

Tenía un trabajo ☐ (1)
No trabajaba ☐ (2)
No contesta ☐ (9)

5. ¿Cuántas horas trabajó? _____ horas
No contesta ☐ (99)
No pertinente (no trabajó) ☐ (97)

6. En relación a las afirmaciones siguientes indique su grado de acuerdo o desacuerdo:

	Totalmente en desacuerdo	*En desacuerdo*	*Ni de acuerdo ni en desacuerdo*	*De acuerdo*	*Totalmente en desacuerdo*	*NS*	*NC*
1. La inmigración es uno de los principales problemas en Europa hoy en día	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)
2. De no controlar las fronteras de Europa, nuestro Estado de Bienestar será insostenible	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)
3. La inmigración ha hecho aumentar la inseguridad en la calle	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)
4. El asentamiento de inmigrantes extracomunitarios está provocando una pérdida de los derechos laborales adquiridos hasta ahora	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)
5. Es necesario implementar políticas de cooperación con los países de origen	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)
6. Los inmigrantes deberían tener derecho a voto	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)
7. Los inmigrantes deben adaptarse a la cultura del país donde se instalan	☐ (1)	☐ (2)	☐ (3)	☐ (4)	☐ (5)	☐ (8)	☐ (9)

7. En política se suele hablar de izquierda y de derecha. En esta tarjeta hay una serie de casillas que van de izquierda a derecha. ¿En qué casilla se colocaría Ud.? MOSTRAR

Izq.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Der.**

No sabe ☐ (98)
No contesta ☐ (99)

Imaginemos el caso del primer cuestionario, una persona que responde:

"Tengo 35 años, soy varón, tengo estudios universitarios terminados, mi padre no tiene los estudios primarios terminados y mi madre bachillerato elemental; estoy trabajando, hago 40 horas semanales. Estoy completamente de acuerdo con que "*La inmigración es uno de los principales problemas en Europa hoy en día*", de acuerdo con que "*De no controlar las fronteras de Europa, nuestro Estado de Bienestar será insostenible*", estoy completamente en desacuerdo con que "*La inmigración ha hecho aumentar la inseguridad en la calle*", en desacuerdo con que "*El asentamiento de inmigrantes extracomunitarios está provocando una pérdida de los derechos laborales adquiridos hasta ahora*", completamente en desacuerdo "*Es necesario implementar políticas de cooperación con los países de origen para que disminuya la entrada de inmigrantes extracomunitarios*" , de acuerdo con que "*Los inmigrantes deberían tener derecho a voto*", completamente de acuerdo con que "*Los inmigrantes han de adaptarse a la cultura del país donde se instalan*". Me sitúo en la casilla 3 entre izquierda y derecha."

La codificación de sus respuestas se recoge en la Tabla III.2.1:

Tabla III.2.1 Codificación de las respuestas del primer individuo de la encuesta

	ID	P1	P2	P3_1	P3_2	P3_3	P4	P5	P6_1	P6_2	P6_3	P6_4	P6_5	P6_6	P6_7	P7
SPSS	1	35	1	6	1	2	1	40	5	4	1	2	1	4	5	3
R	1	35	Varón	Universitario	EGB	Bachillerato	Trabaja	40	CDesacuerdo	Acuerdo	CDesacuerdo	Desacuerdo	CDesacuerdo	Acuerdo	CDesacuerdo	3

Hemos seguido un doble criterio, primero introduciendo solamente códigos numéricos, y después combinando códigos numéricos con texto. El primer caso servirá para la creación e identificación de los datos en SPSS (apartado 1.1) donde se puede codificar toda la información numéricamente y asignar una etiqueta a los códigos cuyo significado requiera ser explicitado, que es el caso de las variables cualitativas. La segunda codificación será la necesaria en R (apartado 1.2) donde se mantienen códigos numéricos para las variables cuantitativas y códigos textuales sintéticos para las variables cualitativas pues en R no es posible diferenciar los valores o códigos de las etiquetas.

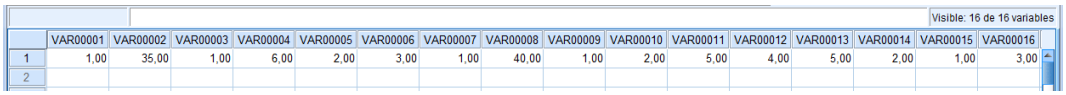
1.1. Creación e identificación de los datos con SPSS

1.1.1. *Introducción de datos en SPSS*

Empezaremos con la tarea de introducción de los datos, más tarde veremos cómo importarlos. Si entramos en la aplicación podemos acceder directamente al editor de datos para introducir la información. Recordemos que si tenemos abierta una matriz de datos previamente en el editor y queremos crear una nueva procederemos en primer lugar a abrir una nueva ventana del **editor de datos en blanco**: Archivo / Nuevo / Datos. El editor de datos permite crear o examinar una matriz de datos a partir de dos pestañas:

En la vista de datos introduciremos los datos propiamente, es decir, los códigos o valores de las variables, mientras que en la vista de variables identificaremos las características de éstos, su diccionario. Podemos optar tanto por empezar a introducir los datos como por elaborar el diccionario.

Procederemos en primer lugar a introducir los datos del primer individuo en el visor de datos de la forma siguiente:



Visible: 16 de 16 variables

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008	VAR00009	VAR00010	VAR00011	VAR00012	VAR00013	VAR00014	VAR00015	VAR00016
1	1,00	35,00	1,00	6,00	2,00	3,00	1,00	40,00	1,00	2,00	5,00	4,00	5,00	2,00	1,00	3,00
2																

Automáticamente se genera un nombre para cada variable y se les asigna el formato por defecto: tipo numérico de anchura 8 y 2 decimales, sin etiquetas, valores perdidos ni nivel de medición. La imagen inicial de la pestaña de variables es la siguiente:

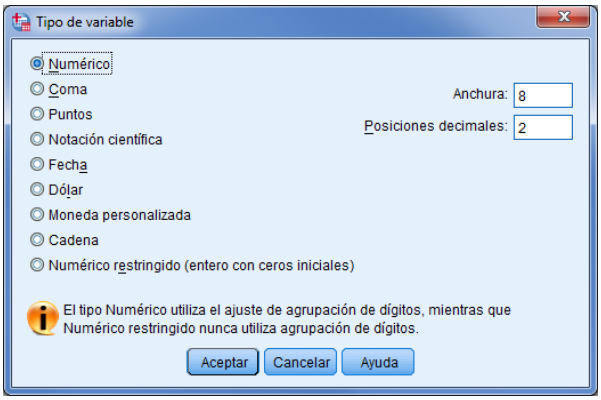
	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	VAR00001	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
2	VAR00002	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
3	VAR00003	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
4	VAR00004	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
5	VAR00005	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
6	VAR00006	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
7	VAR00007	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
8	VAR00008	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
9	VAR00009	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
10	VAR00010	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
11	VAR00011	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
12	VAR00012	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
13	VAR00013	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
14	VAR00014	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
15	VAR00015	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada
16	VAR00016	Numérico	8	2		Ninguna	Ninguna	7	Derecha	Desconocido	Entrada

Procedemos a continuación a identificar y crear el diccionario de los datos en la vista de variables, lo que implica especificar la información siguiente en relación a cada variable que se disponen en las filas:

- El **nombre de la variable** (Nombre): puede tener una extensión de 64 caracteres, deben empezar con una letra del alfabeto (**A-Z**) o con los signos **@**, y también **#** para una variable temporal y **$** para una variable del sistema; el resto puede ser además número, un "**.**" o un "**_**". Pero no pueden acabar en punto, ni valen los espacios o caracteres especiales como **!**, **?**, **'** o *****. Es indiferente utilizar mayúsculas o minúsculas, conservándose la forma elegida. Las palabras clave ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO y WITH no se pueden utilizar.

- El **tipo de formato** de la variable (Tipo): cada variable es un tipo de dato que se define según los tipos siguientes: **numérico** (los valores son números en formato estándar), **coma** y **punto** (tipo numérico que acepta la coma o el punto como separador cada tres posiciones), **notación científica** (numérico cuyos valores se muestran con una E intercalada y un exponente con signo que representa una potencia de base 10), **fecha** (variable numérica con diferentes formatos fecha-calendario u hora-reloj) **dólar** o **moneda personalizada** (variable numérica que se muestra con un signo dólar inicial ($) o en los formatos definidos en opciones),

cadena (los valores son textuales con cualquier carácter) y **numérico restringido** (valores enteros no negativos)[5]:



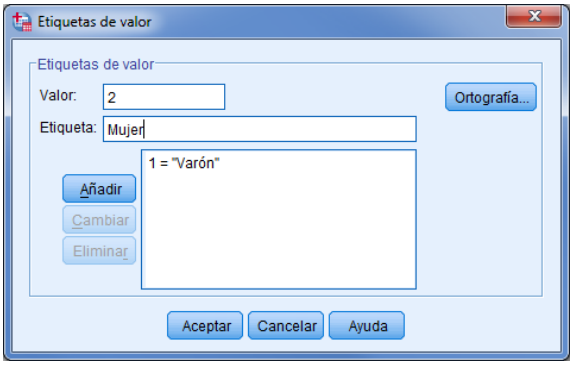
- Las posiciones (Anchura) son los dígitos que ocupa la variable, una parte de los cuales corresponde al número de decimales (Decimales). Se puede precisar tanto en el cuadro de diálogo del tipo como en su propia columna. En el caso de las variables de tipo cadena, fecha y numérico restringido el número de decimales siempre es 0.
  Se recomienda utilizar en general el **formato numérico estándar** pues facilita el tratamiento de las variables. Con un mayor dominio del software o para necesidades específicas evidentemente todos los formatos son válidos. El formato numérico estándar se define por defecto con el formato F8.2, es decir, con 8 posiciones de anchura y 2 decimales que se corresponden con la siguiente disposición: 5 posiciones del número entero, una posición para el punto decimal y 2 posiciones de los decimales: _ _ _ _ _ . _ _ . Así, por ejemplo, el valor 1 de la variable número de hijos se corresponde con 00001.00 y es visualizado como 1.00. Si cambiamos la variable a formato F1.0 será entonces simplemente 1. En cualquiera de los dos casos no afecta más que a la forma de verse.

- La etiqueta de la variable (Etiqueta) permite asignar un texto identificativo del contenido de la misma, con una extensión máxima de 256 caracteres. No obstante en muchos resultados no es posible ver la etiqueta en toda su extensión. En general 36 caracteres pueden ser suficientes[6]. La etiqueta se escribe directamente sobre la casilla.

- Las etiquetas de los valores de las variables (Valores) asignan un texto identificativo de su significado, con una extensión máxima de 120 caracteres, pero con 16 caracteres como máximo puede ser suficiente. Para consignarla se clica sobre el lado derecho de la casilla y se accede a un cuadro de diálogo donde se escribe cada valor con su etiqueta y se clica sobre "Añadir":

[5] Para acceder al cuadro de diálogo para definir el tipo de variable es necesario clicar sobre el lado derecho de la casilla: Numérico.

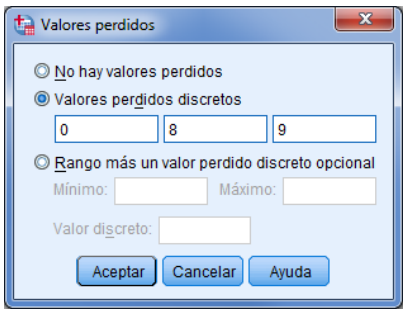
[6] En las etiquetas de las variables y de los valores se pueden insertar los símbolos \n para forzar un salto de línea.



- Los valores perdidos declarados por el usuario (Perdidos). Es habitual que nos encontramos en la situación de ausencia de valores, de no disponer de información para algunos casos o individuos en relación a una o más variables. El sistema necesita, sin embargo, identificar igualmente estas situaciones con un valor determinado. Estos valores se denominan valores perdidos (*missing values*). Los hay de dos tipos:

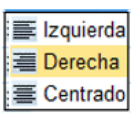
  - **Valores perdidos del usuario**. Son los valores que implican una falta de información (por ejemplo, los "no sabe", "no contesta" o "no pertinente") que se codifican con un valor determinado (por ejemplo, 8, 9 y 0), y se declaran por el usuario como perdidos en la identificación de las variables para tratar de forma diferenciada y que, por defecto, no forman parte de los cálculos.
  - **Valores perdidos del sistema**. Se corresponden también con la falta de información, pero se generan automáticamente por el software cuando encuentran una casilla en blanco en la matriz de datos, o bien cuando generamos una nueva variable y no se asigna un valor determinado a uno o más casos. Los valores perdidos se visualizan en el editor con un punto (".") Y en las tablas aparecen con la etiqueta "Perdidos Sistema".

  Los valores perdidos del usuario son los que se identifican en el diccionario de los datos. Para ello es necesario clicar sobre el lado derecho de la casilla y se accede al cuadro de diálogo donde se detallan valores concretos (hasta 3) o rangos de valores:



- El ancho de columna que se visualiza en el editor de datos (Columnas).

- Se puede controlar la presentación de los valores de los datos y/o de las etiquetas de valor en la vista de datos (Alineación): izquierda, derecha y centrado.



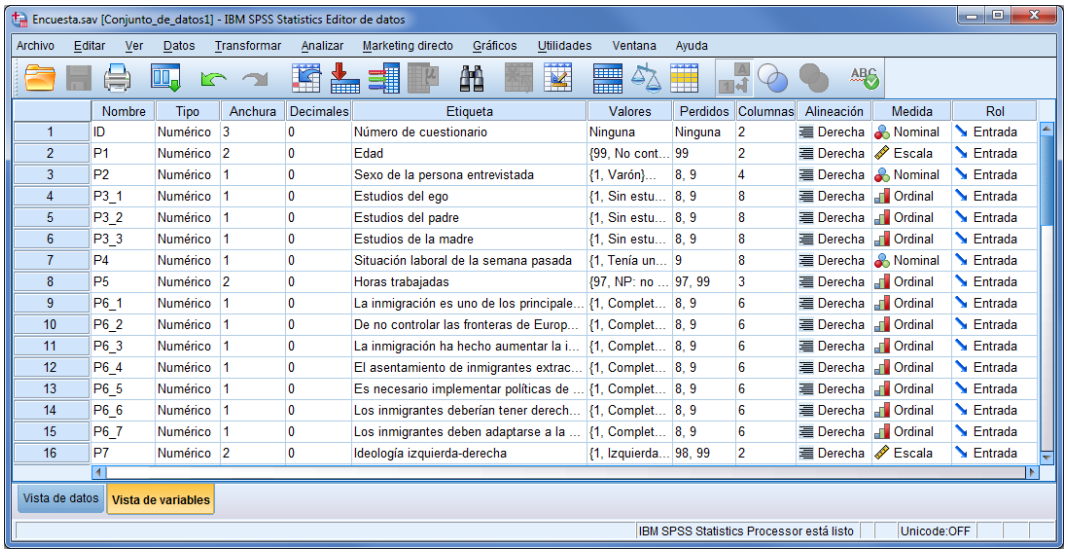
- El nivel de medida de cada variable (Medida) por defecto es desconocido y es conveniente definirlo pues en algunos procedimientos se tiene en cuenta para decidir el tipo de análisis o de gráfico. En otros casos, la mayor parte, el procedimiento acepta cualquier nivel de medida; como usuarios debemos ser conocedores de qué escala de medición de las variables se emplea en cada caso. En SPSS se diferencian tres niveles de medición:

- El papel de la variable (Rol) identifica un tipo particular de variable con una función específica que se predefine y permite preseleccionar variables para el análisis sólo en los cuadros de diálogo. Los roles disponibles son: **entrada** (la variable se utiliza como independiente, opción por defecto), **salida** (variables resultado o dependiente), ambos (doble papel de entrada y salida), **ninguno** (sin función), **partición** (variable que sirve para segmentar los datos) y **dividir** (para compatibilidad con *IBM SPSS Modeler*).

Cada uno de los atributos que definen el diccionario de cada variable se puede copiar y pegarlo a continuación en la definición de otra (u otras) variable(s). También se pueden copiar (y borrar) variables enteras seleccionando una línea[7].

Con estas indicaciones procedemos a realizar la identificación de los datos con las propiedades particulares de cada una de las variables. El resultado final aparece en el Gráfico III.2.2.

Gráfico III.2.2 Identificación de los datos de la encuesta: vista de variables



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	ID	Numérico	3	0	Número de cuestionario	Ninguna	Ninguna	2	Derecha	Nominal	Entrada
2	P1	Numérico	2	0	Edad	{99, No cont...	99	2	Derecha	Escala	Entrada
3	P2	Numérico	1	0	Sexo de la persona entrevistada	{1, Varón}...	8, 9	4	Derecha	Nominal	Entrada
4	P3_1	Numérico	1	0	Estudios del ego	{1, Sin estu...	8, 9	8	Derecha	Ordinal	Entrada
5	P3_2	Numérico	1	0	Estudios del padre	{1, Sin estu...	8, 9	8	Derecha	Ordinal	Entrada
6	P3_3	Numérico	1	0	Estudios de la madre	{1, Sin estu...	8, 9	8	Derecha	Ordinal	Entrada
7	P4	Numérico	1	0	Situación laboral de la semana pasada	{1, Tenía un...	9	8	Derecha	Nominal	Entrada
8	P5	Numérico	2	0	Horas trabajadas	{97, NP: no ...	97, 99	3	Derecha	Ordinal	Entrada
9	P6_1	Numérico	1	0	La inmigración es uno de los principale...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
10	P6_2	Numérico	1	0	De no controlar las fronteras de Europ...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
11	P6_3	Numérico	1	0	La inmigración ha hecho aumentar la i...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
12	P6_4	Numérico	1	0	El asentamiento de inmigrantes extrac...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
13	P6_5	Numérico	1	0	Es necesario implementar políticas de ...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
14	P6_6	Numérico	1	0	Los inmigrantes deberían tener derech...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
15	P6_7	Numérico	1	0	Los inmigrantes deben adaptarse a la ...	{1, Complet...	8, 9	6	Derecha	Ordinal	Entrada
16	P7	Numérico	2	0	Ideología izquierda-derecha	{1, Izquierda...	98, 99	2	Derecha	Escala	Entrada

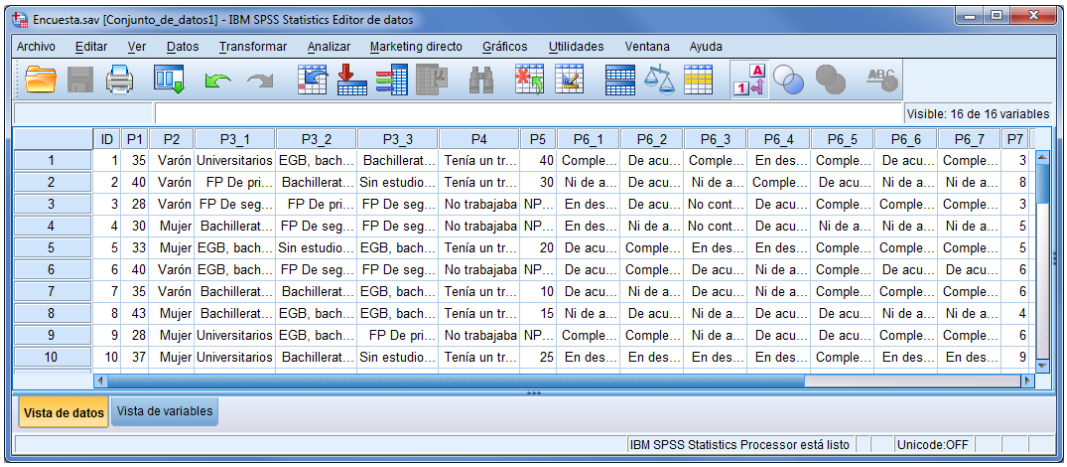
La identificación realizada desde la ventana del editor de datos también se puede elaborar con el lenguaje de comando de SPSS. El archivo de sintaxis Encuestas.sps incluye esta información.

[7] Las columnas de los atributos se pueden reorganizar, para ello es necesario ir al menú: Ver / Personalizar vista de variables. También se puede crear atributos personalizados desde el menú: Datos / Nuevo atributo personalizado.

Definido el diccionario o las propiedades de las variables nos queda completar la información de la matriz de datos con la introducción de los valores en la vista de datos[8]. En nuestro caso hemos introducido 9 casos más que dan lugar a una imagen como la del Gráfico III.2.3.

La introducción de los datos no implica más que colocarse sobre la casilla correspondiente e introducir el valor de los datos y darle a <*Enter*> o ir a otra casilla. Cuando se introducen los datos de variables cualitativas, si no tenemos realizada la precodificación de los datos aparte y por tanto tenemos que elegir el código, debemos consultarlo. Para ello disponemos de una opción de visualización bastante interesante en SPSS. Es necesario activar primero el botón  de Etiquetas de valor. A continuación, sobre la casilla que queremos introducir el valor, se clica sobre el lado derecho de la casilla donde estemos, se abrirá un desplegable donde podremos elegir con el botón derecho del ratón y elegir la etiqueta que corresponde al valor. Esta opción de visualización de las etiquetas de los valores es igualmente interesante en una análisis habitual de los datos pues las variables que aparecen con etiquetas son las cualitativas o categóricas (nominales y ordinales), mientras que en las cuantitativas el valor numérico ya habla por sí solo y no requiere una etiqueta identificativa.

Gráfico III.2.3 Identificación de los datos de la encuesta: vista de datos



En el editor de datos se puede:

- **Insertar** filas (casos) o columnas (variables) mediante la selección primero de una fila o de una columna para determinar el punto de inserción y, a continuación, a través del menú contextual clicar sobre Insertar caso o Insertar variable. Estas acciones se pueden ejecutar a través del menú "Datos" o bien a través de los iconos de la barra de herramientas:  .
- **Eliminar** filas (casos) o columnas (variables) mediante la selección de la fila o de la columna (o más de una), y pulsamos sobre <SUPR> o hacemos <CTRL>+<X> (o con el menú Edición o con el menú contextual).
- **Copiar** filas (casos) o columnas (variables) con <CTRL>+<C> o con el menú "Edición" o con el menú contextual.

[8] A la vista de datos se puede ir desde la vista de variables haciendo doble clic sobre una fila de variable en la vista de variables. De forma equivalente, desde la vista de datos se puede ir a la vista de variables haciendo doble clic sobre el nombre de la columna de una variable en la vista de datos.

- **Pegar** filas (casos) o columnas (variables) con <Ctrl>+<V> o con el menú "Edición" o con el menú contextual.
- Podemos **deshacer** o **rehacer** acciones a través de los iconos  .
- **Buscar** valores a través del icono de la barra de herramientas:  o a través del menú "Edición".

Una vez introducidos los datos, o a medida que los vamos grabando para no perder el trabajo realizado, debemos guardarlos y convertirlos en un fichero del sistema SPSS, por ejemplo con el nombre Encuesta.sav[9]. Para **guardar un archivo de datos**:

- A través del menú: Archivo / Guardar o bien Archivo / Guardar como
- Con el teclado: Ctrl+S
- Clicando sobre el botón "Guardar este documento"  .

Una vez creada la matriz de datos podemos pedirle al SPSS la información del diccionario de los datos. A través del menú: Archivo / Mostrar información del archivo de datos, eligiendo archivo de trabajo, pues se puede elegir entre éste (el que esté abierto en el editor) o de otro archivo externo que esté guardado en el disco (Gráfico III.2.4). Este procedimiento corresponde con el comando de sintaxis del SPSS: **DISPLAY DICTIONARY**.

Gráfico III.2.4 Listado del diccionario de los datos de la matriz de datos de la encuesta

Información de variable

Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación	Valores perdidos
ID	1	Número de cuestionario	Nominal	Entrada	2	Derecha	F3	F3	
P1	2	Edad	Escala	Entrada	2	Derecha	F2	F2	99
P2	3	Sexo de la persona entrevistada	Nominal	Entrada	4	Derecha	F1	F1	8, 9
P3_1	4	Estudios del ego	Ordinal	Entrada	8	Derecha	F1	F1	8, 9
P3_2	5	Estudios del padre	Ordinal	Entrada	8	Derecha	F1	F1	8, 9
P3_3	6	Estudios de la madre	Ordinal	Entrada	8	Derecha	F1	F1	8, 9
P4	7	Situación laboral de la semana pasada	Nominal	Entrada	8	Derecha	F1	F1	9
P5	8	Horas trabajadas	Ordinal	Entrada	3	Derecha	F2	F2	97, 99
P6_1	9	La inmigración es uno de los principales problemas en Europa hoy en día	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P6_2	10	De no controlar las fronteras de Europa, nuestro Estado de Bienestar será insostenible	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P6_3	11	La inmigración ha hecho aumentar la inseguridad en la calle	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P6_4	12	El asentamiento de inmigrantes extracomunitarios está provocando una pérdida de los derechos laborales adquiridos hasta ahora	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P6_5	13	Es necesario implementar políticas de cooperación con los países de origen	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P6_6	14	Los inmigrantes deberían tener derecho a voto	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P6_7	15	Los inmigrantes deben adaptarse a la cultura del país donde se instalanlen	Ordinal	Entrada	6	Derecha	F1	F1	8, 9
P7	16	Ideología izquierda-derecha	Escala	Entrada	2	Derecha	F2	F2	98, 99

Variables en el archivo de trabajo

[9] Esta matriz de datos se encuentra en la página web del capítulo.

. . .

Valores de variable

Valor		Etiqueta
P1	99[a]	No contesta
P2	1	Varón
	2	Mujer
P3_1	1	Sin estudios, primarios inacabados
	2	EGB, bachillerato elemental, ESO
	3	Bachillerato superior, BUP, COU
	4	FP De primer grado, oficialías
	5	FP De segundo grado, maestría industrial
	6	Universitarios
	8[a]	No sabe
	9[a]	No contesta
P3_2	1	Sin estudios, primarios inacabados
	2	EGB, bachillerato elemental, ESO
	3	Bachillerato superior, BUP, COU
	4	FP De primer grado, oficialías
	5	FP De segundo grado, maestría industrial
	6	Universitarios
	8[a]	No sabe
	9[a]	No contesta
P3_3	1	Sin estudios, primarios inacabados
	2	EGB, bachillerato elemental, ESO
	3	Bachillerato superior, BUP, COU
	4	FP De primer grado, oficialías
	5	FP De segundo grado, maestría industrial
	6	Universitarios
	8[a]	No sabe
	9[a]	No contesta
P4	1	Tenía un trabajo
	2	No trabajaba
	9[a]	No contesta
P5	97[a]	NP: no trabajó
	99[a]	No contesta
P6_1	1	Completamente de acuerdo
	2	De acuerdo
	3	Ni de acuerdo ni en desacuerdo
	4	En desacuerdo
	5	Completamente en desacuerdo
	9[a]	No contesta
P6_2	1	Completamente de acuerdo
	2	De acuerdo
	3	Ni de acuerdo ni en desacuerdo
	4	En desacuerdo
	5	Completamente en desacuerdo
	9[a]	N
• • •		
P6_7	1	Completamente de acuerdo
	2	De acuerdo
	3	Ni de acuerdo ni en desacuerdo
	4	En desacuerdo
	5	Completamente en desacuerdo
	9[a]	No contesta
P7	1	Izquierda
	10	Derecha
	98[a]	No sabe
	99[a]	No contesta

a. Valor perdido

Asimismo el procedimiento Libro de Códigos (comando **CODEBOOK** del SPSS) que se ejecuta en el menú: Analizar / Informes / Libro de Códigos, permite obtener la información del diccionario y los estadísticos de resumen de las variables especificadas que elijamos: recuentos y porcentajes con variables nominales y ordinales; y media, desviación típica y cuartiles para las variables de escala.

Gráfico III.2.5 Libro de códigos de algunas variables de la matriz de la encuesta

P2

		Valor	Recuento	Porcentaje
Atributos estándar	Posición	3		
	Etiqueta	Sexo de la persona entrevistada		
	Tipo	Numérico		
	Formato	F1		
	Medición	Nominal		
	Rol	Entrada		
Valores válidos	1	Varón	5	50,0%
	2	Mujer	5	50,0%

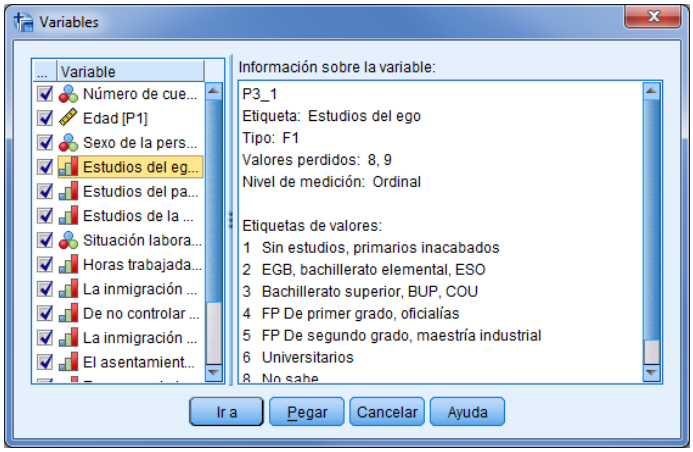
P3_1

		Valor	Recuento	Porcentaje
Atributos estándar	Posición	4		
	Etiqueta	Estudios del ego		
	Tipo	Numérico		
	Formato	F1		
	Medición	Ordinal		
	Rol	Entrada		
Valores válidos	1	Sin estudios, primarios inacabados	0	0,0%
	2	EGB, bachillerato elemental, ESO	2	20,0%
	3	Bachillerato superior, BUP, COU	3	30,0%
	4	FP De primer grado, oficialías	1	10,0%
	5	FP De segundo grado, maestría industrial	1	10,0%
	6	Universitarios	3	30,0%
Valores perdidos	8	No sabe	0	0,0%
	9	No contesta	0	0,0%

P7

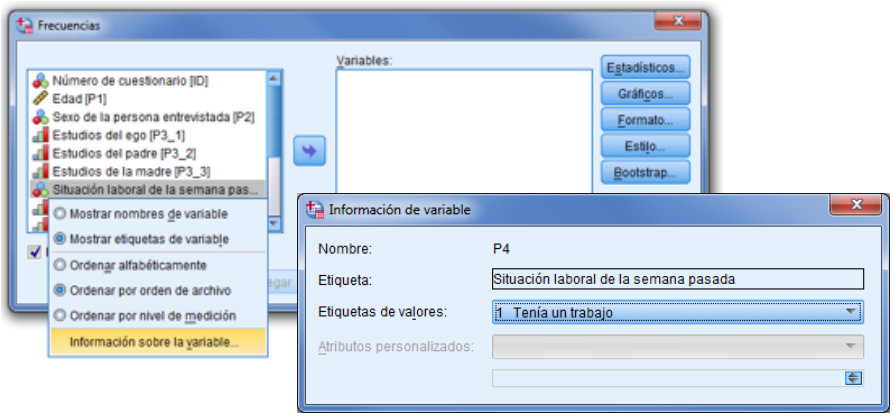
		Valor	Recuento	Porcentaje
Atributos estándar	Posición	16		
	Etiqueta	Ideología izquierda-derecha		
	Tipo	Numérico		
	Formato	F2		
	Medición	Escala		
	Rol	Entrada		
N	Válido	10		
	Perdidos	0		
Tendencia y dispersión centrales	Media	5,50		
	Desviación estándar	1,958		
	Percentil 25	4,00		
	Percentil 50	5,50		
	Percentil 75	6,00		
Valores etiquetados	1	Izquierda	0	0,0%
	10	Derecha	0	0,0%
	98	No sabe	0	0,0%
	99	No contesta	0	0,0%

El diccionario de los datos, además de poder consultarlo en la pestaña de vista de variables y de los procedimientos anteriores, se puede consultar en cualquier momento a través a del icono Variables  de la barra de herramientas. Cuando se clica se obtiene un cuadro como el siguiente:



donde se informa de las principales propiedades de cada variable.

Finalmente la información de una variable también se puede consultar dentro de un cuadro de diálogo de un menú pulsando con el botón derecho sobre la variable y a continuación haciendo clic sobre Información sobre la variable. Por ejemplo desde el menú de Frecuencias:



Una vez identificados los datos, un modo de comprobar la corrección del trabajo realizado es pedir las tablas de frecuencias a través del menú Analizar / Estadísticos descriptivos / Frecuencias. Seleccionamos las variables y las pasamos al recuadro de Variables pulsando sobre el icono  . Finalmente ejecutamos el procedimiento de obtener las frecuencias pulsando sobre Aceptar.

Finalmente solo comentar que el diccionario de una variable se puede aplicar a otras a través del menú Datos / Copiar propiedades de datos (comando **APPLY DICTIONARY** de SPSS), ya sea desde un archivo de datos externo o desde un conjunto de datos abierto.

► Ejercicio 1. Propuesto
A partir de la matriz de datos creada Encuesta.sav obtener las tablas de frecuencias de las distintas variables y comprobar la correcta identificación de los datos.

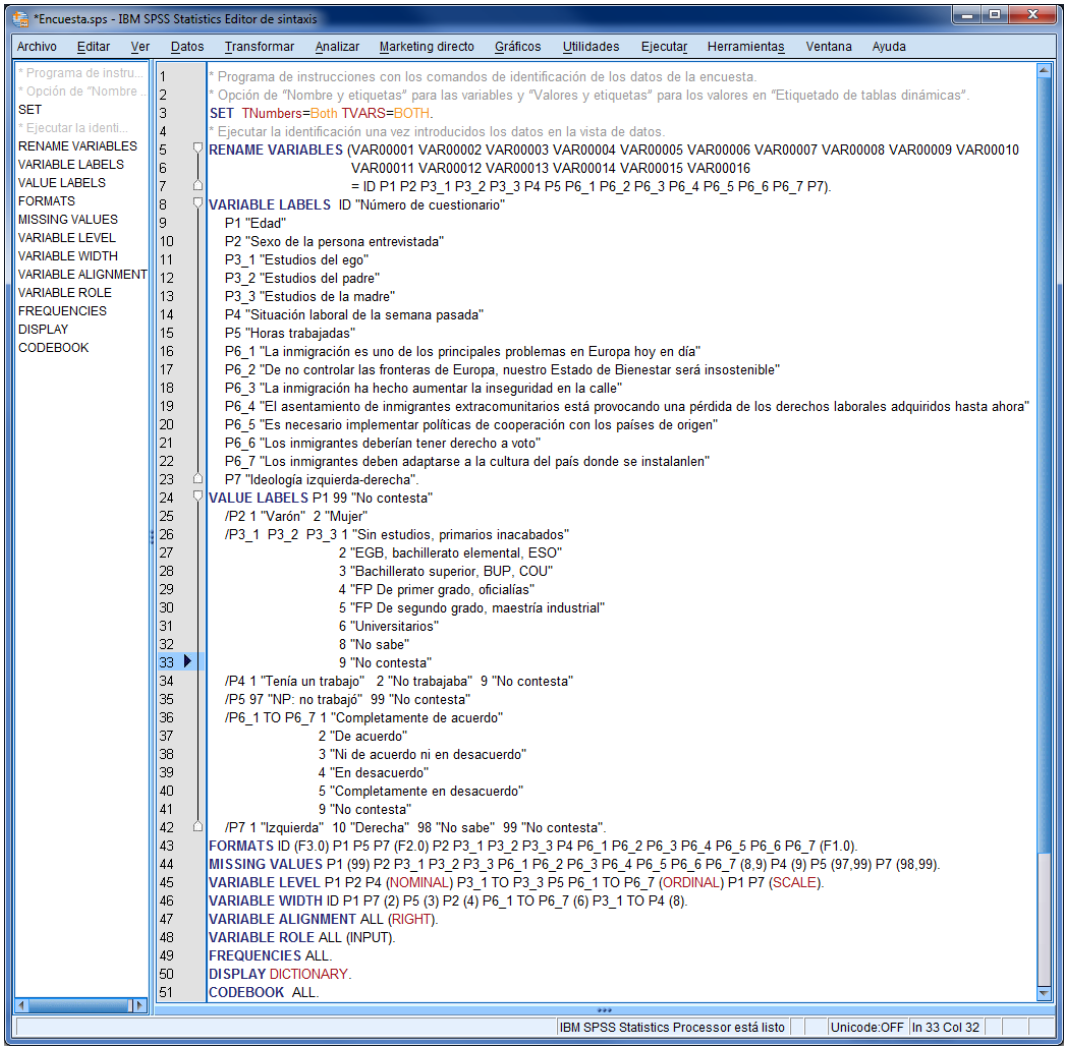
► Ejercicio 2. Propuesto
Con la matriz de datos CIS3041.sav obtener el diccionario de los datos y el libro de códigos para las variables: CCAA, TAMUNI, P3, P901, P1001, P1101, P1301, P15, P1601, P1701, P18, P2013, P23, P25, P28, P29, P31, P32, P46, VOTOSIM, RECUERDO, ESTUDIOS, OCUMAR11, CONDICION y ESTATUS, que permiten reconocer los principales tipos de variables y preguntas del Barómetro del CIS.
También se pueden pedir las tablas de frecuencias de todas ellas.
Recordemos el interés de tener activadas las opciones "Nombre y etiquetas" para las variables y "Valores y etiquetas" para los valores en "Etiquetado de tablas dinámicas".

Para finalizar este apartado se adjunta en el Gráfico III.2.6 la imagen del archivo de sintaxis que realiza los distintos aspectos de identificación que hemos ido comentando. En el archivo Encuesta.sps de la página web se encuentra dicha sintaxis. Comentamos brevemente la sintaxis utilizada.

Al inicio se introducen unos comentarios que se indican en la sintaxis iniciando el texto del comentario con un asterisco (*). Antes de proceder a la identificación se activan las opciones que comentamos en el capítulo anterior de activación de nombres y etiquetas de las variables y valores y etiquetas de los valores de las variables.

Si introducimos primero los datos sin nombrar a las variables el sistema SPSS hemos visto que le asigna un nombre por defecto. El comando **RENAME VARIABLES** cambia el nombre original por el que hemos acordado.

Gráfico III.2.6 Sintaxis para la identificación de los datos de la encuesta. Encuesta.sps

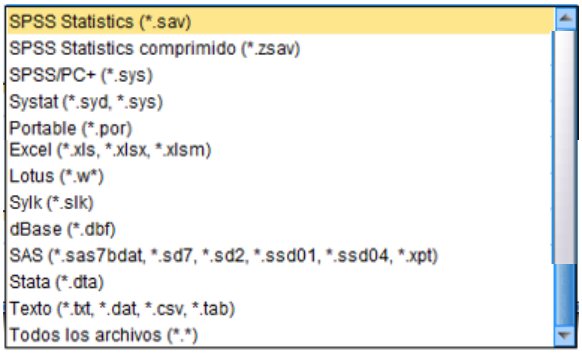


A continuación se asignan etiquetas a las variables (comando **VARIABLE LABELS**) y también a los valores de las variables (comando **VALUE LABELS**). El comando **FORMATS** determina el tipo de formato de las variables, en nuestro caso todas las variables son numéricas y se definen con tres anchuras diferentes y sin decimales: F1.0, F2.0 y F3.0. Los valores perdidos se especifican con el comando **MISSING VALUES** especificando entre paréntesis después de cada grupo de variables los valores que el usuario define como perdidos. El nivel de medición se fija con el comando **VARIABLE LEVEL**: agrupamos las variables en tres bloques y asigmamos entre paréntesis los tres niveles posibles. Con **VARIABLE WIDTH** se especifica el ancho de la columna en el

editor de datos y con **VARIABLE ALIGNMENT** la alineación de los valores de las casillas. Por último la identificación del diccionario se completa con el rol que se asigna a las variables (comando **VARIABLE ROLE**). Se completa el programa de sintaxis con tres instrucciones más destinadas a obtener las tablas de frecuencias de todas las variables (comando **FREQUENCIES**), a listar el diccionario de las variables que hemos creado (comando **DISPLAY DICTIONARY**) y el libro de códigos (comando **CODEBOOK**).

1.1.2. *Importación y exportación de datos en SPSS*

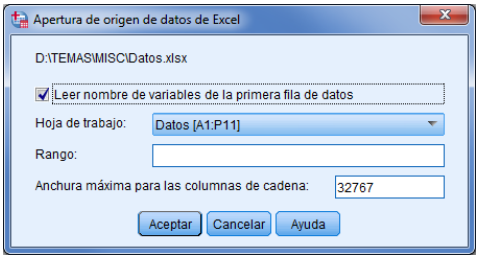
Los archivos de datos creados en otro software con un formato definido (SPSS, SAS, Excel,…) o bien sin formato, de texto plano (DAT, TXT), se puede importar fácilmente desde SPSS. A través del menú Archivo / Abrir / Datos de SPSS o con las teclas <CTRL>+<O>, o el botón del editor de datos, accedemos a un cuadro de diálogo que nos permite abrir un fichero eligiendo entre una diversidad de formatos:



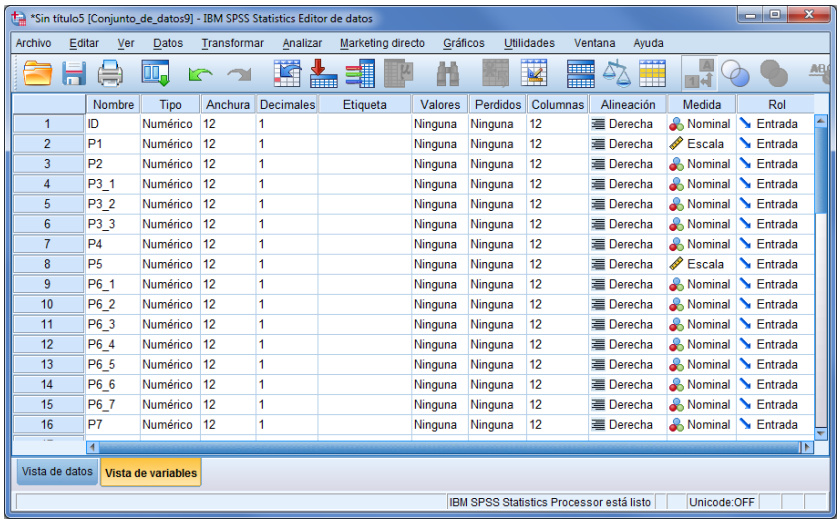
Por un lado están tres formatos propios de SPSS, además del habitual sav: uno que comprime los datos (zsav), otro que abre un formato antiguo de la versión del software que se llamó SPSS/PC+ (sys) y el formato portable que permite llevarlo entre sistemas operativos donde está instalado el SPSS. El resto de los formatos hacen referencia a otros paquetes estadísticos como Systat, SAS o Stata, a hojas de cálculo como Excel, Lotus o Sylk, a gestores de bases de datos como dBase, además formatos de texto plano, es decir, sin formato, donde los datos están separados por comas, tabulaciones, espacios,… (txt, dat, csv, tab).

En la página web de este capítulo se encuentran los archivos Datos.xlsx, Datos.csv y Datos.dat, que utilizaremos para realizar un ejercicio de importación. Se pueden importar directamente abriéndolos y completando los cuadros de diálogo que aparecerá. En todos los casos se trata de la matriz de datos que hemos identificado más arriba y guardado como Encuesta.sav, con toda la información codificada numéricamente.

En el caso de abrir o importar el archivo de Excel Datos.xlsx, aparece un cuadro de diálogo para definir la hoja de datos, el rango de los datos y para informar de la existencia de una primera línea con el nombre de las variables:

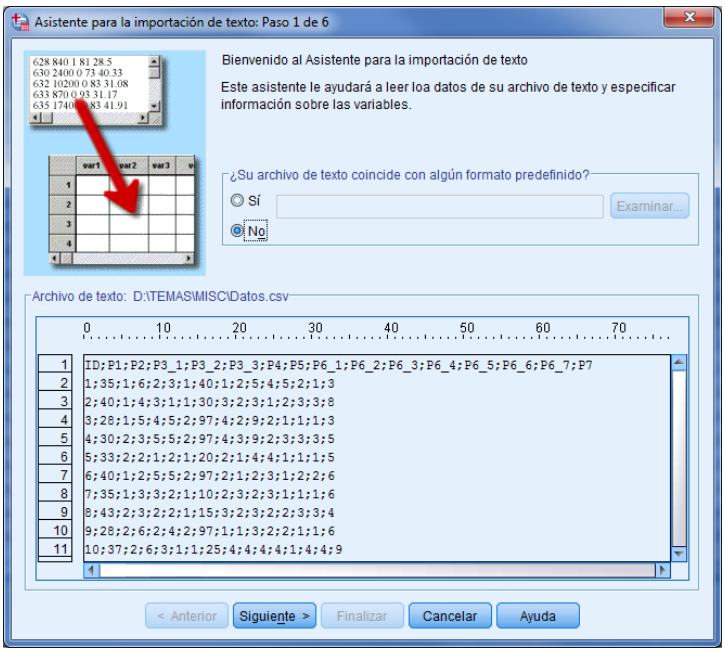


Después de aceptar aparecen los datos en el editor con los nombres de las variables y el formato numérico para todas ellas. Por tanto, será necesario completar el diccionario de los datos con toda la información de etiquetas, valores perdidos y demás formatos.

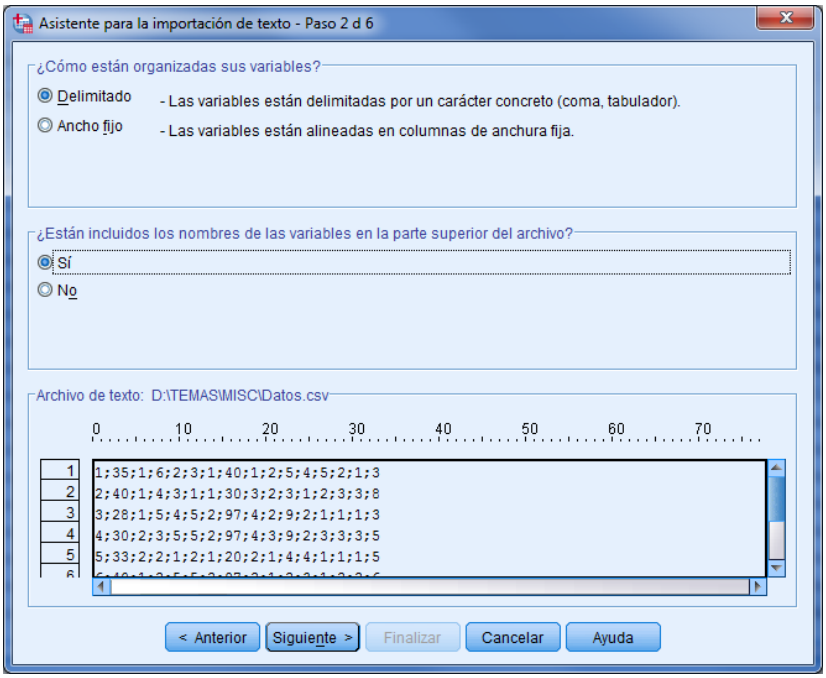


	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	ID	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
2	P1	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Escala	Entrada
3	P2	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
4	P3_1	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
5	P3_2	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
6	P3_3	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
7	P4	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
8	P5	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Escala	Entrada
9	P6_1	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
10	P6_2	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
11	P6_3	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
12	P6_4	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
13	P6_5	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
14	P6_6	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
15	P6_7	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada
16	P7	Numérico	12	1		Ninguna	Ninguna	12	Derecha	Nominal	Entrada

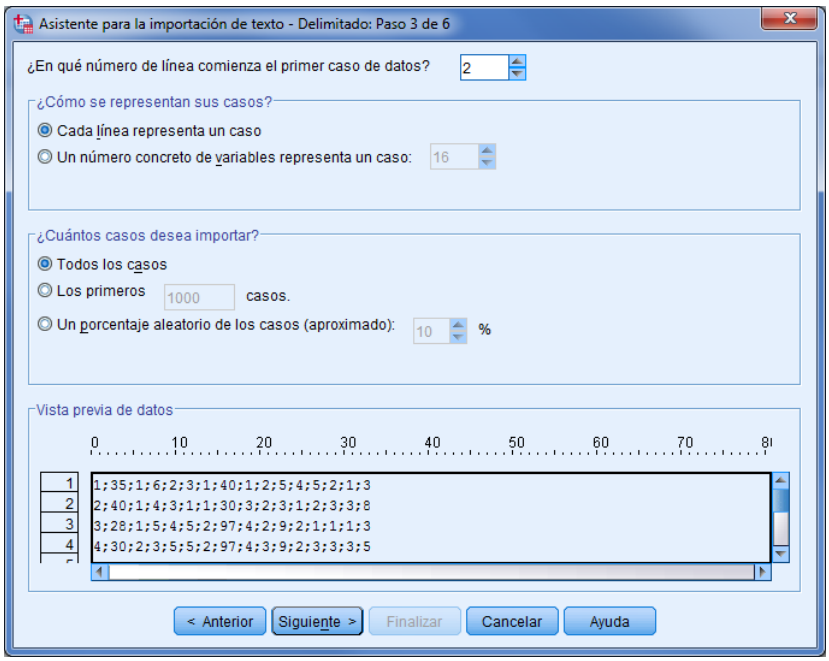
En el caso de los archivos Datos.csv y Datos.dat ambos se corresponden con un formato de datos delimitados, por punto y coma en el primer caso y por tabulaciones en el segundo. El proceso de importación es similar, lo veremos con el primero de los archivos. Una vez se abre aparece este cuadro de diálogo, el primero de seis:



En él se visualiza la disposición de los datos y se determina si se corresponde con algún formato que tengamos predefinido. Clicamos sobre siguiente y nos aparece el segundo cuadro de diálogo:

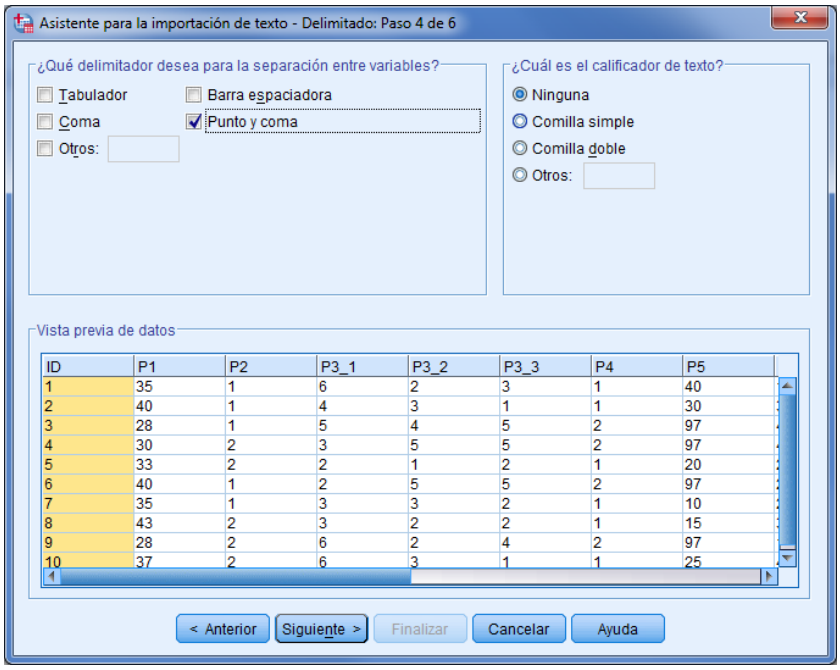


Se determina si los datos están delimitados, como es nuestro caso, o bien si los datos se disponen alineados en columnas con una anchura determinada[10]. También se informa de si el nombre de las variables aparece en la primera fila del archivo. Pasamos a la siguiente ventana:

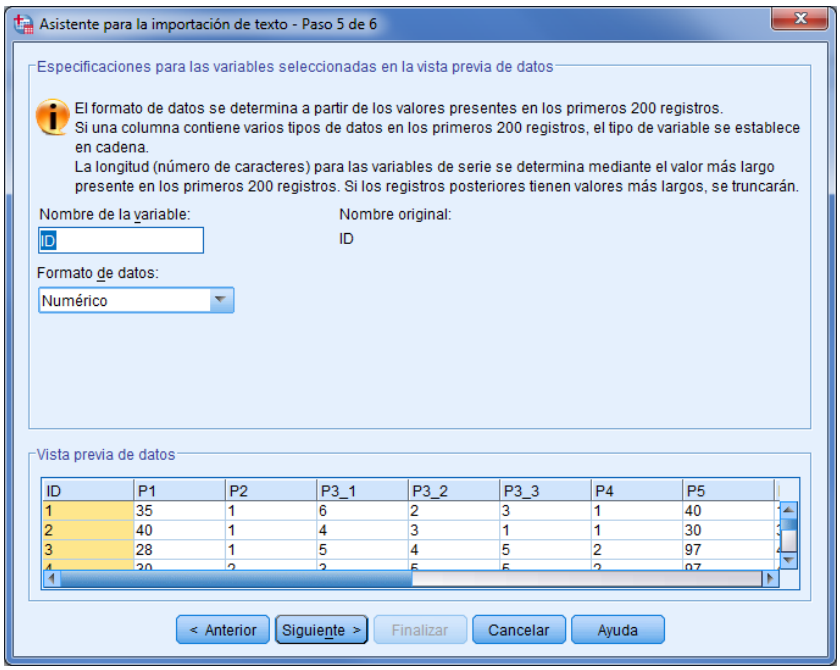


[10] Más adelante (apartado 1.1.3) presentaremos el ejemplo de importación e identificación de los datos del Barómetro y otras encuestas del CIS cuyos datos que se presentan en formato de texto con una disposición fija de columna.

En este caso configuramos la importación indicando que los datos empiezan en la fila 2, que cada registro (fila) corresponde a un caso y que importe todos los casos. Pasamos a la cuarta ventana:

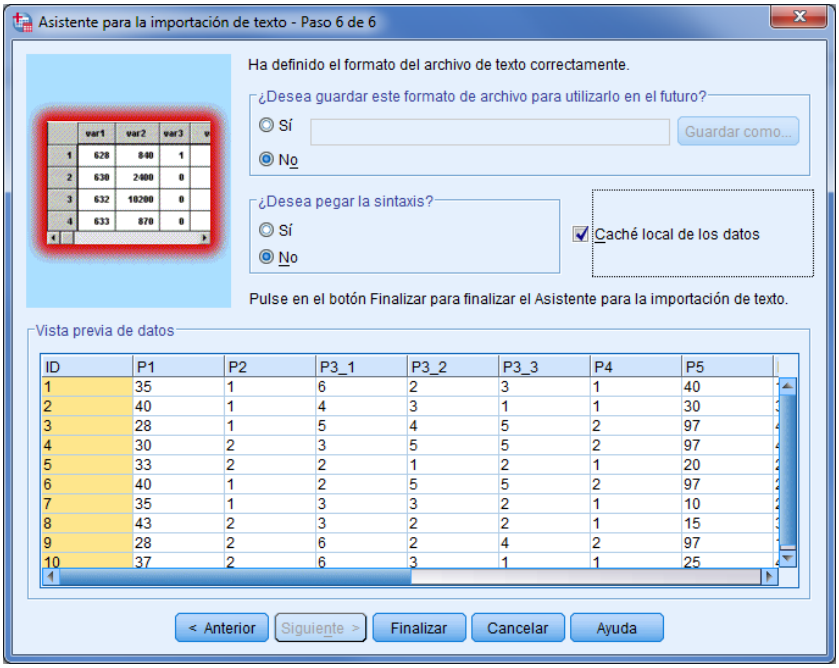


Aquí especificamos el delimitador, en nuestro caso el punto y coma, y si tenemos datos textuales que estén delimitados entre caracteres particulares. Seguidamente en el quinto paso:



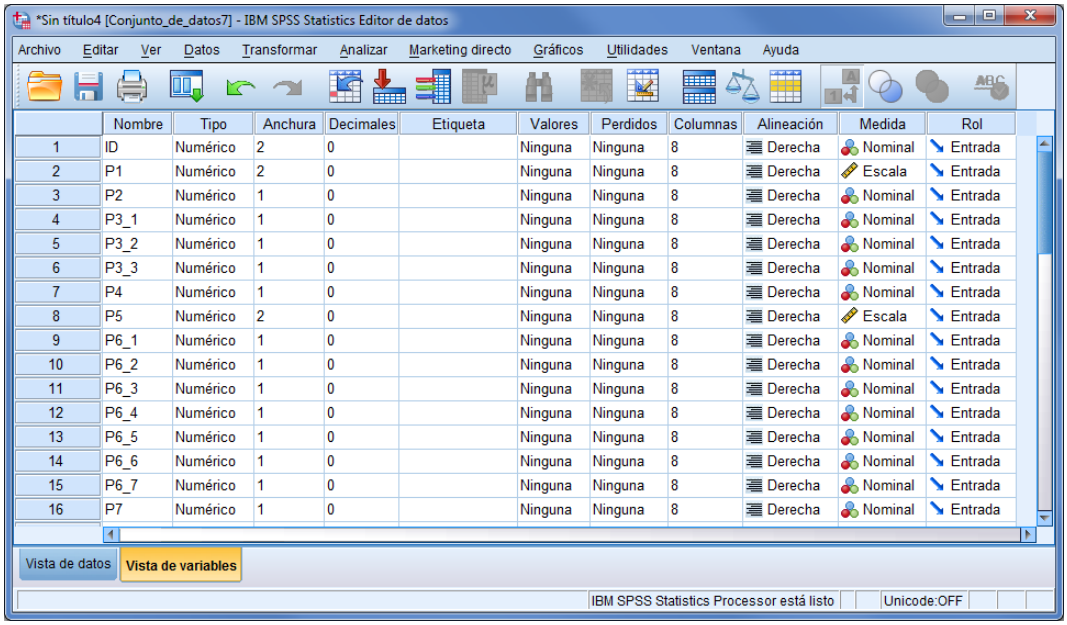
podemos cambiar el nombre a las variables y el tipo de formato de los datos de cada una de las variables (numérico, cadena,…).

Finalmente llegamos a la sexta y última etapa:

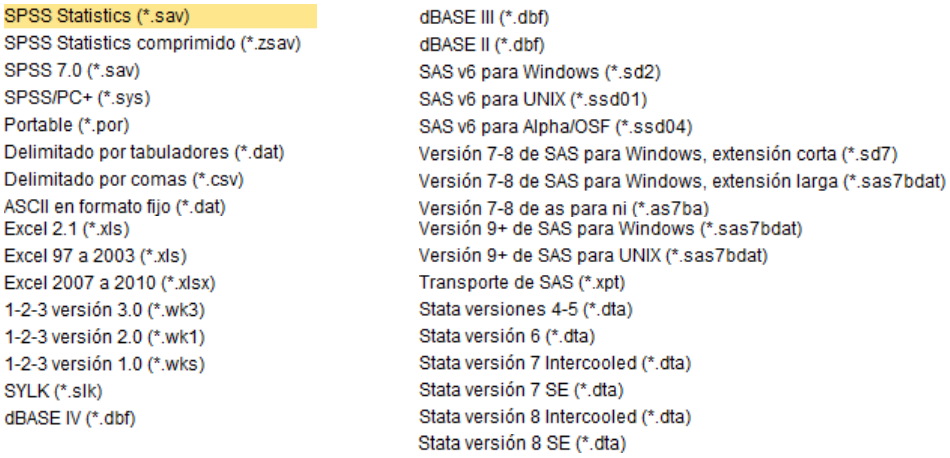


En este último cuadro del asistente podemos guardar el formato utilizado para otra ocasión y optar por la ejecución inmediata de la importación o convertir esa acción en lenguaje de comandos del SPSS que se adjuntará en una ventana de sintaxis. Para acabar clicamos sobre Finalizar.

Como en el caso anterior solamente hemos importado los datos, los nombres de las variables y se han definido parte de los formatos posibles. El resto como etiquetas o valores perdidos hay que completarlos seguidamente.



Por otra parte podemos tener la necesidad de exportar nuestros datos desde SPSS hacia otras aplicaciones. También podemos guardar (exportar) nuestros datos en diferentes formatos. Cuando hacemos Guardar o Guardar como tenemos disponibles estas alternativas en el desplegable Guardar como tipo:



1.1.3. *Importación e identificación de los datos de las encuestas del CIS*

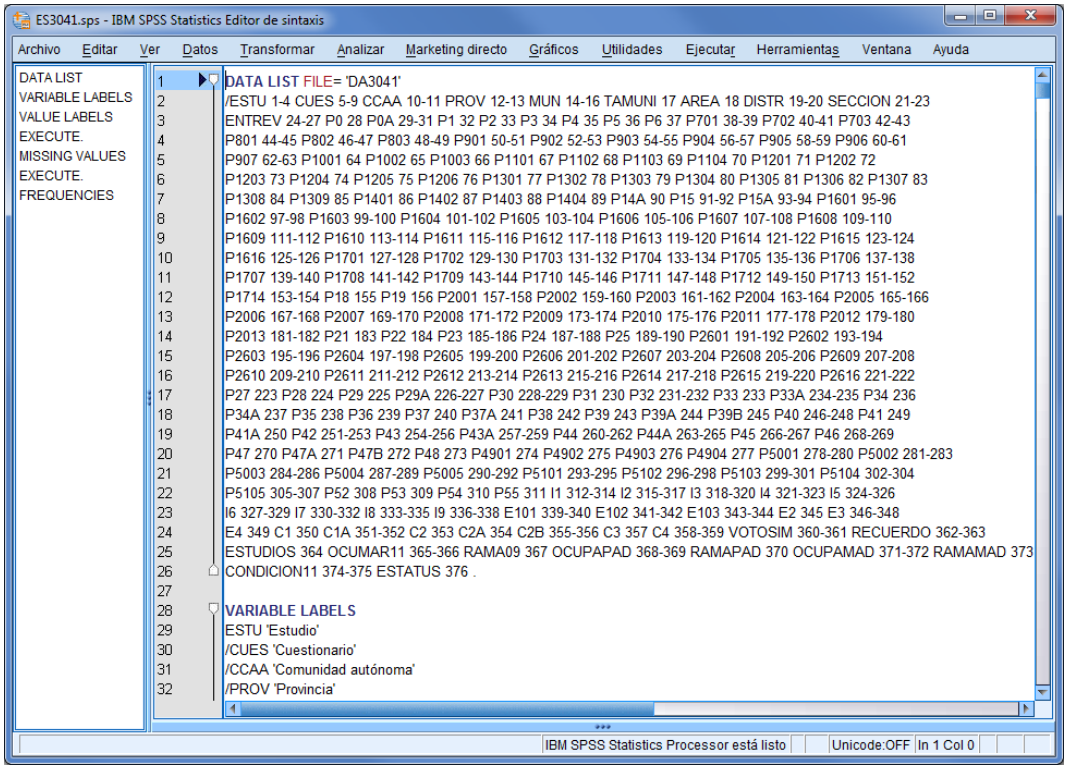
Desde el 1 de enero de 2009 el Centro de Investigaciones Sociológicas (http://www.cis.es/)[11] pone a disposición de las personas interesadas, de forma gratuita, los ficheros de datos de las encuestas realizadas por este organismo. Los ficheros de datos tienen formato ASCII (formato plano, TXT o DAT) y se pueden descargar desde la página web del CIS así como los archivos de sintaxis de los paquetes estadísticos SPSS y SAS, el cuestionario, la ficha técnica, el libro de códigos y las tarjetas, a través de la dirección: http://www.cis.es/cis/opencms/CA/2_bancodatos/. En este manual manejamos esta fuente de información que consideramos esencial para el conocimiento de la realidad política y social española, además de constituir un recurso docente valiosísimo en la enseñanza y aprendizaje de la metodología de la investigación cuantitativa. Por ello es de interés conocer con mayor detalle el procedimiento de importación e identificación de los datos del CIS en SPSS. Lo haremos además presentando el lenguaje de sintaxis que ejecuta esta tarea.

Una vez bajado el archivo de los datos de interés (MDxxxx.zip), en nuestro caso nos referiremos al estudio número 3041 correspondiente al Barómetro del mes de octubre de 2014, es necesario descomprimirlo y seleccionar dos de los archivos que incluye el fichero zip. Por un lado el archivo DA con el número contiene los datos sin formato. Se puede abrir con el Bloc de notas o con Excel y ver su contenido. Por otro lado el archivo SE con el número corresponde al archivo de sintaxis del SPSS. Se puede cambiar su nombre ESnº por ESnº.sps para abrirlo directamente con el software SPSS y ejecutar la sintaxis.

[11] El Centro de Investigaciones Sociológicas (CIS) es un organismo autónomo dependiente del Ministerio de la Presidencia de España, con la función principal de contribuir al conocimiento científico de la sociedad española.

En la página web de este capítulo se puede encontrar el archivo ES3041.sps que proporciona el CIS y que parcialmente se reproduce en el Gráfico III.2.7. El programa de instrucciones se puede seleccionar y ejecutar teniendo la precaución de ubicar el archivo de datos DA3041 en la misma carpeta de trabajo del software.

Gráfico III.2.7 Archivos de sintaxis del CIS para la identificación de los datos



Alternativamente tenemos dos opciones para asegurar que se localizarán los datos. Por un lado podemos hacer uso del comando **CD** (cambiar de directorio) que indica al sistema cuál es la carpeta de trabajo por defecto (por ejemplo C:\Datos), colocándola en la primera línea de archivo de sintaxis:

```
CD 'C:\Datos'.
```

Por otro, podemos especificar la ruta del archivo en el comando **DATA LIST**:

```
DATA LIST FILE 'C:\Datos\DA3041'.
```

Finalmente se selecciona todo, se ejecuta y se guarda el archivo de datos que se genera, en nuestro caso lo guardamos con el nombre CIS3041.sav.

Los datos del CIS se disponen en un formato fijo de columna, es decir, cada variable se ubica en unas columnas específicas que afectan a todos los individuos y alinean verticalmente todos los datos. Las columnas que ocupa cada variable vienen especificadas en el cuestionario por un número entre paréntesis al lado derecho de las categorías de respuesta y en el libro de códigos.

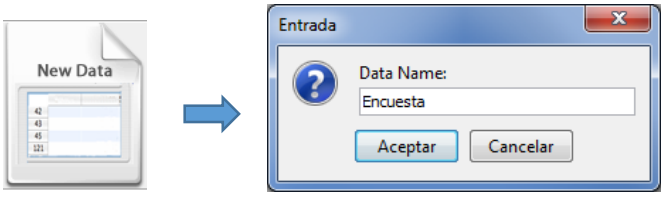
Para identificar esta información podríamos utilizar el procedimiento anterior aplicado a Datos.csv con el asistente de importación. Mediante la sintaxis que comentamos se emplea el comando **DATA LIST**, destinado a definir los datos adaptándose a su disposición y asignando nombre y tipo de formato. En el formato fijo de columna se coloca el nombre de cada variable y los números de las columnas que ocupa. Adicionalmente se le puede asignar el tipo de formato (tipo, anchura y decimales), en este caso la anchura viene dada por las columnas que ocupa cada variable y se asigna por defecto formato numérico a todas las variables. Si tuviéramos decimales o la variable tuviera un formato distinto se precisaría detallarlo en el comando.

El programa de sintaxis se completa asignado etiquetas a las variables (comando **VARIABLE LABELS**), etiquetas a los valores (comando **VALUE LABELS**), asignando los valores perdidos (comando **MISSING VALUES**) y pidiendo las tablas de frecuencias de todas las variables (comando **FREQUENCIES**)[12].

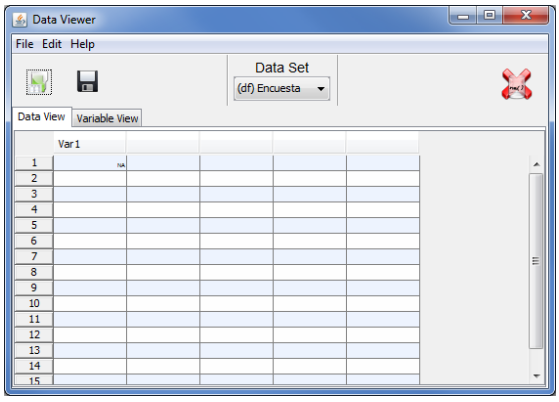
1.2. Creación e identificación de los datos con R

1.2.1. *Introducción de datos en R*

Nuestra primera tarea será de introducción de los datos y más tarde veremos cómo importarlos en R. Realizaremos esta tarea con Deducer que nos facilitará el trabajo de creación e identificación en un entorno de ventanas. Para crea una matriz de datos, si acabamos de entrar en Deducer, tendremos la opción de clicar sobre New Data en la ventana inicial de *Data Viewer*, nos aparecerá seguidamente un cuadro para darle un nombre que no contenga ni acentos ni espacios. Le podremos el nombre de Encuesta:



Se abrirá el **editor de datos en blanco**:

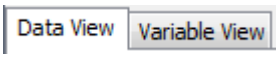


[12] En la matriz de datos CIS3041.sav hemos incorporado una identificación más completa de los datos pues algunas variables no son identificadas con etiquetas de variables y de valores, por otro lado la definición de valores perdidos se puede ampliar para considerar también las respuestas de "no sabe" y "no contesta", y también se ha definido el nivel de medición de las variables.

Si estuviéramos trabajando con otros datos, desde el editor abierto procederemos a abrir una nueva ventana del editor de datos en blanco mediante: File / New Data / Datos, o bien con las teclas <CTRL>+<N>.

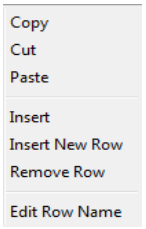
El tipo de archivos de datos con los que trabajaremos habitualmente, nuestras matrices de datos, se identifican en R como *data frames*.

El editor de datos, que abre al inicio o a partir del menú de la consola de se permite crear o examinar una matriz de datos a partir de dos pestañas:

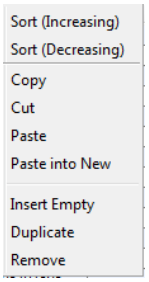


En la *Data View* introduciremos los datos propiamente, es decir, los códigos o valores de las variables, mientras que en la *Variable View* identificaremos las características de éstos, su diccionario. Podríamos optar tanto por empezar a introducir los datos como por elaborar el diccionario, pero es recomendable proceder en primer lugar a introducir los datos, pues nos ayudarán, en el caso de las variables cualitativas, a generar automáticamente el diccionario de sus valores.

En el visor de datos si clicamos el botón derecho del ratón en cualquier fila, además de copiar, cortar y pegar, podremos: insertar una nueva fila (*Insert New Row*), borrarla (*Remove Row*) y cambiar el nombre de la fila (*Edit Row Name*). Desde el momento que creamos una nueva línea ésta aparece con el valor NA (*Not Available*) en cada casilla que identifica la ausencia de valor (casilla en blanco).



Si clicamos el botón derecho del ratón en cualquier columna, además de copiar, cortar y pegar, podremos: insertar una nueva columna vacía (*Insert Empty*), borrarla (*Remove*), o duplicarla (*Duplicate*), así como ordenar los datos de la columna de forma ascendente o descendente (*Sort: Increasing-Decreasing*).



Consideremos las respuestas del primer individuo que sugerimos en la Tabla III.2.1: 1, 35, Varón, Universitarios, EGB, Bachillerato, Trabaja, 40, CDesacuerdo, Acuerdo, CDesacuerdo, Desacuerdo, CDesacuerdo, Acuerdo, CDesacuerdo, 3.

y las introduciremos literalmente en el visor de datos, en la fila 1, de la forma siguiente:

	Var1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	1	35	Varón	Universitarios	EGB	Bachillerato	Trabaja	40	CDesacuerdo	Acuerdo	CDesacuerdo	Desacuerdo	CDesacuerdo	Acuerdo	CDesacuerdo	3
2																

Automáticamente se genera un nombre para cada variable y se les asigna el formato por defecto según el valor que hemos introducido. Si nos situamos en el visor de variables la imagen inicial de la pestaña de *Variable View* es la siguiente:

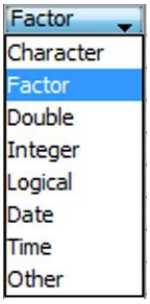
	Variable	Type	Factor Levels
1	Var1	Double	
2	V2	Double	
3	V3	Character	
4	V4	Character	
5	V5	Character	
6	V6	Character	
7	V7	Character	
8	V8	Double	
9	V9	Character	
10	V10	Character	
11	V11	Character	
12	V12	Character	
13	V13	Character	
14	V14	Character	
15	V15	Character	
16	V16	Double	

Los valores que hemos introducido con códigos numéricos tienen el formato *Double* mientras que los valores con código textual se identifican con el formato *Character*.

Procedemos a continuación a modificar esta información de identificación y creación del diccionario de los datos. Para ello especificaremos la información siguiente en relación a cada variable que se dispone en las filas:

- El **nombre de la variable** (Variable): tiene que empezar con una letra o con punto, el nombre que se asigna es distinto si se escribe con mayúsculas o minúsculas, no pueden tener acentos, ni ñ ni ç, ni espacios en blanco, ni ningún carácter fuera del estándar inglés, tampoco admite los símbolos de los operadores aritméticos.

- El **tipo de formato** de la variable (Type): las variables de un *data frame* de R pueden ser de diferente tipo. En particular podemos hacer la distinción fundamental entre:
  - **Cualitativas o categóricas**: valores de texto o etiqueta (numérica o textual) que representa el grupo o categoría a la que pertenece el caso. Se pueden diferenciar entre nominales (por ejemplo el sexo) y ordinales (nivel de estudios). En R se denominan factores, y en el caso de ser de nivel ordinal factores ordenados.
  - **Cuantitativas**: valores numéricos con los que tiene sentido realizar aritmética. Se pueden diferenciar entre continuas (índice de masa corporal) y discretas (número de hijos). En R se llaman double si tienen decimales e integer si representan datos discretos.

  Cuando clicamos sobre cada casilla de la columna *Type* se abre un desplegable que nos permite definir el formato de la variable.

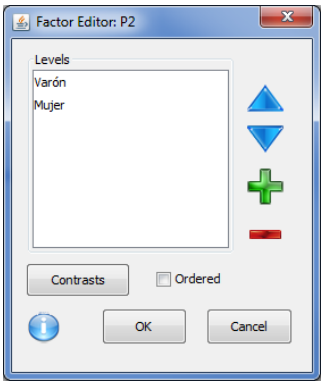
  Así, el paquete estadístico Deducer clasifica los tipos de variables en:
  - *Character*: variables cadena (texto).
  - *Factor*: variables categóricas que pueden ser nominales u ordinales.
  - *Double*: variables cuantitativas continuas.
  - *Integer*: variables cuantitativas discretas.
  - *Logical*: variables lógicas o dicotómicas.
  - *Date*: variables de fechas.
  - *Time*: variables de tiempo.
  - Otros tipos de variables

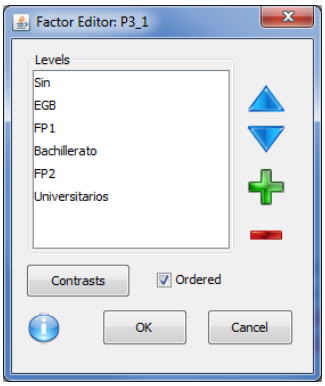
- Los **valores de las variables factor** (Factor Levels): se detallan las etiquetas o valores de estas variables que tratamos como cualitativas, de nivel de medida

nominal u ordinal, y donde hay que especificar cada etiqueta o valor de la variable. Las etiquetas se pueden definir y editar clicando en la propia celda.

Cuando creamos una matriz de datos no es necesario definir de antemano las etiquetas de la variable factor. Como veremos, a medida que se introducen los datos las siguientes etiquetas se irán incorporando automáticamente.

Cada etiqueta o valor de las variables cualitativas que se introduce es un texto que identifica a cada categoría de la variable, y el conjunto de las etiquetas se ordenan según el orden de introducción: o bien en el editor del factor o bien en la vista de datos. Este orden puede ser relevante para las características de la variable y puede resultar que la introducción de las etiquetas no se adecúe a lo que queremos. Con las flechas  las podemos ordenar. También podemos añadirlas con  o quitarlas con  .





Cuando en particular la categoría de la variable (*level*) pueda tomar varios valores ordenables siguiendo una escala preestablecida (variable ordinal) marcaremos la casilla Ordered. También se pueden modificar a través de la consola en el menú Data / Edit Factor.

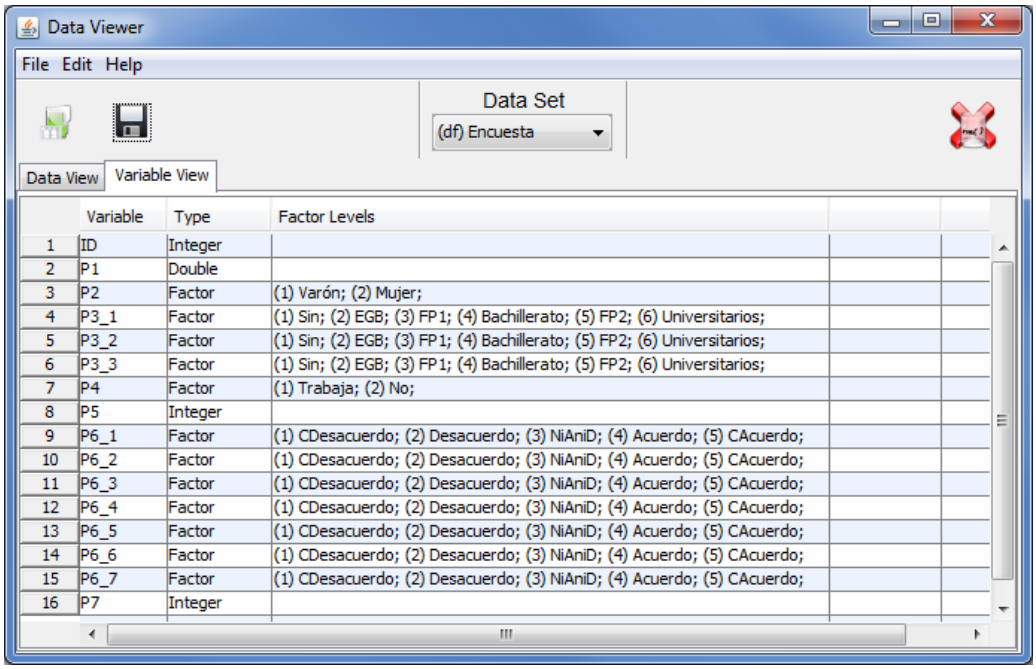
Finalmente hay que destacar que cada etiqueta se identifica en el sistema de R con un valor entero consecutivo que aparece entre paréntesis en cada celda de la variable precisando numéricamente el orden.

Un aspecto importante en la identificación y en el posterior tratamiento y análisis de los datos es la ausencia de valores, los llamados valores perdidos (*missing values*). Es habitual que nos encontremos en la situación de no disponer de información de algunos casos o individuos en relación a una o más variables, por ejemplo, en los casos no sabe, no contesta o no pertinente, se trata de información que habitualmente no se procesa, por lo tanto, para poder realizar de forma correcta los análisis y su interpretación se deben tratar de forma específica. A diferencia de otros paquetes estadísticos donde se pueden asignar valores específicos a cada situación y tratarlos de formas distinta, en R la solución es drástica: cualquier valor que sea considerado perdido no se codifica y se tratan de forma unificada identificándolos con el símbolo NA (*Not Available*). En R no es necesario asignarles ningún valor particular,

sencillamente se corresponden con un "agujero" de información de la matriz, casillas que se dejan en blanco y que reconocemos porque aparecen la letras NA[13].

Con estos criterios procedemos a realizar la identificación de los datos con las propiedades particulares de cada una de las variables. El resultado final del diccionario de datos aparece en el Gráfico III.2.8 y los datos se pueden visualizar en el Gráfico III.2.9. Para llegar a ese resultado primero hemos cambiado el nombre de las variables, hemos precisado a continuación su tipo y finalmente hemos codificado los datos de las variables factor. Para la codificación se pueden utilizar los códigos disponibles en la imagen de la pestaña del visor de variables que ilustra el Gráfico III.2.9[14].

Gráfico III.2.8 Identificación de los datos de la encuesta: vista de variables



	Variable	Type	Factor Levels
1	ID	Integer	
2	P1	Double	
3	P2	Factor	(1) Varón; (2) Mujer;
4	P3_1	Factor	(1) Sin; (2) EGB; (3) FP1; (4) Bachillerato; (5) FP2; (6) Universitarios;
5	P3_2	Factor	(1) Sin; (2) EGB; (3) FP1; (4) Bachillerato; (5) FP2; (6) Universitarios;
6	P3_3	Factor	(1) Sin; (2) EGB; (3) FP1; (4) Bachillerato; (5) FP2; (6) Universitarios;
7	P4	Factor	(1) Trabaja; (2) No;
8	P5	Integer	
9	P6_1	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
10	P6_2	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
11	P6_3	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
12	P6_4	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
13	P6_5	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
14	P6_6	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
15	P6_7	Factor	(1) CDesacuerdo; (2) Desacuerdo; (3) NiAniD; (4) Acuerdo; (5) CAcuerdo;
16	P7	Integer	

Los valores o categorías de las variables cualitativas no hay que introducirlos necesariamente desde el visor de variables, el sistema los puede crear automáticamente a medida que introducimos los datos en la pestaña del visor de datos, además les asigna internamente un valor numérico que indica el orden de posición de cada categoría de la variable. En el momento de introducir los datos en el *Data View*, Deducer interpreta el tipo de variable según la información proporcionada, e incluso cambia (sin aviso) el tipo. Esto puede generar problemas: si definimos una variable como *integer* pero introducimos un número con decimales, 2.0 por ejemplo, la convierte en *double*, si introducimos un número decimal con coma 2,3 en vez de punto 2.3 la convierte en *character*. En R, y por tanto en Deducer, el separador de decimales es el punto, y no la

[13] Una forma alternativa de tratar en R de forma diferencial estos valores perdidos es (1) codificarlos con un valor diferenciado, (2) crear una copia de la variable original en la cual los valores perdidos correspondientes estén en blanco (NA), y (3) realizar los análisis seleccionando la versión de la variable que más interese en cada caso, con o sin NA, o combinando la información de ambas.

[14] En el caso de las variables factor seguimos como criterio utilizar un código sintético de una sola palabra, pudiendo utilizar los acentos. No obstante, trabajar con acentos en R es problemático y obliga a renunciar a la especificidad de la lengua propia en favor de la anglosajona, aspecto que debería ser revisado. En el caso de las variables hemos tomado con criterio de asignación del nombre el número de la pregunta del cuestionario, pero se puede seguir también el criterio de utilizar un nombre sintético que remita a su contenido.

coma. Un dato introducido que contenga una coma no es tratada como numérica, sino como texto.

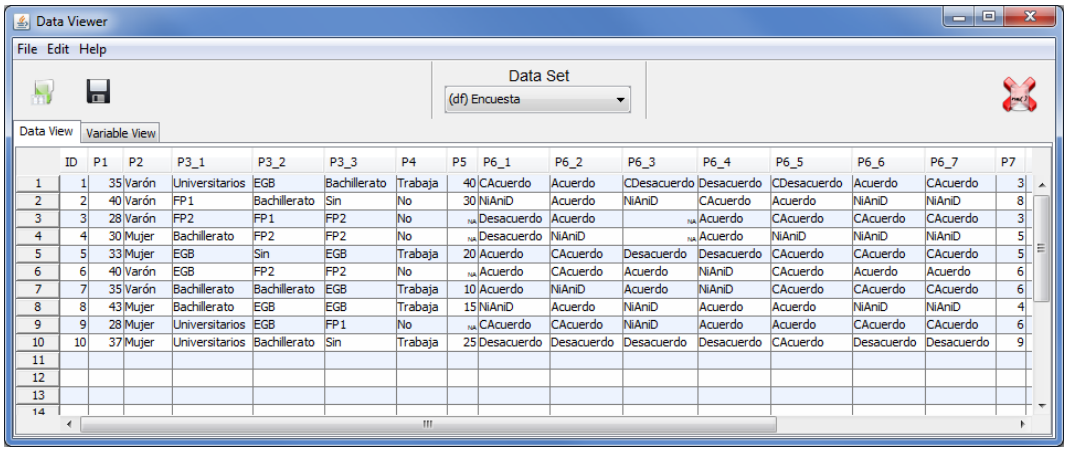
Hay que tener también presente que cada valor (llamado nivel, *level*) de una variable cualitativa (que será de tipo *factor*), será cada conjunto de caracteres diferentes introducidos. Por ejemplo, si escribimos Mujer como valor de la variable Sexo para un individuo y mujer para otro, se considerarán diferentes y tendremos 2 códigos para identificar a las mujeres.

Si la variable factor está medida a nivel ordinal (*ordered factor*) el orden de las categorías es importante cuando se visualiza la información. Este orden de los valores de las variables, cuando los códigos se generan automáticamente a medida que los introducimos en la matriz, no respetan el orden deseado y requiere que editemos los niveles del factor para ordenarlos según el sentido de cada variable.

Por otra parte, hay que tener en cuenta que si editamos los *Factores levels* de una variable cualitativa y borramos por error uno de los niveles, borraremos los datos correspondientes de la matriz y se convertirán en NA (valores perdidos).

En el caso de las respuestas correspondientes a valores perdidos hemos seguido el criterio de considerar las categorías "nos sabe", "no contesta" y "no pertinente" conjuntamente y no asignarles un código específico, por lo que aparecen sin distinción con el símbolo NA en la matriz de datos.

Gráfico III.2.9 Identificación de los datos de la encuesta: vista de datos



	ID	P1	P2	P3_1	P3_2	P3_3	P4	P5	P6_1	P6_2	P6_3	P6_4	P6_5	P6_6	P6_7	P7
1	1	35	Varón	Universitarios	EGB	Bachillerato	Trabaja	40	CAcuerdo	Acuerdo	CDesacuerdo	Desacuerdo	CDesacuerdo	Acuerdo	CAcuerdo	3
2	2	40	Varón	FP1	Bachillerato	Sin	No	30	NiAniD	Acuerdo	NiAniD	CAcuerdo	Acuerdo	NiAniD	NiAniD	8
3	3	28	Varón	FP2	FP1	FP2	No	NA	Desacuerdo	Acuerdo	NA	Acuerdo	CAcuerdo	CAcuerdo	CAcuerdo	3
4	4	30	Mujer	Bachillerato	FP2	FP2	No	NA	Desacuerdo	NiAniD	NA	Acuerdo	NiAniD	NiAniD	NiAniD	5
5	5	33	Mujer	EGB	Sin	EGB	Trabaja	20	Acuerdo	CAcuerdo	Desacuerdo	Desacuerdo	CAcuerdo	CAcuerdo	CAcuerdo	5
6	6	40	Varón	EGB	FP2	FP2	No	NA	Acuerdo	CAcuerdo	Acuerdo	NiAniD	CAcuerdo	Acuerdo	Acuerdo	6
7	7	35	Varón	Bachillerato	Bachillerato	EGB	Trabaja	10	Acuerdo	NiAniD	Acuerdo	NiAniD	CAcuerdo	CAcuerdo	CAcuerdo	6
8	8	43	Mujer	Bachillerato	EGB	EGB	Trabaja	15	NiAniD	Acuerdo	NiAniD	Acuerdo	Acuerdo	NiAniD	NiAniD	4
9	9	28	Mujer	Universitarios	EGB	FP1	No	NA	CAcuerdo	CAcuerdo	NiAniD	Acuerdo	Acuerdo	CAcuerdo	CAcuerdo	6
10	10	37	Mujer	Universitarios	Bachillerato	Sin	Trabaja	25	Desacuerdo	Desacuerdo	Desacuerdo	Desacuerdo	CAcuerdo	Desacuerdo	Desacuerdo	9
11																
12																
13																
14																

Así pues, primero introducimos los datos en el *Data View* como aparece en el Gráfico III.2.9 donde se han grabado 10 casos. La introducción de los datos no implica más que colocarse sobre la casilla correspondiente e introducir el valor de los datos y darle a <Intro> o ir a otra casilla. A continuación modificamos el nombre de las variables, definimos sus tipos y en el caso de las variables factor ajustamos el orden de las categorías y determinamos si son ordinales.

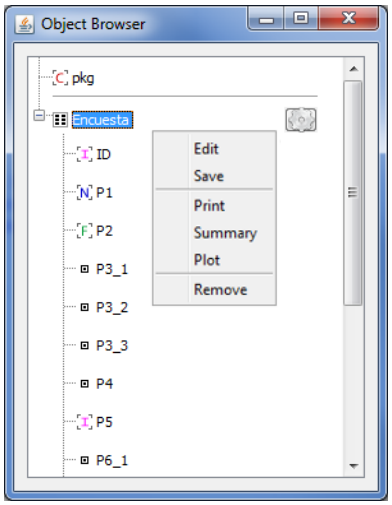
En el editor de datos se puede:

- **Copiar** filas (casos) o columnas (variables) con <CTRL>+<C> o con el menú Edit o con el menú contextual.

- **Cortar** filas (casos) o columnas (variables con <CTRL>+<X> o con el menú Edit o con el menú contextual. No elimina la fila / columna.
- **Pegar** filas (casos) o columnas (variables) con <CTRL>+<V> o con el menú Edit o con el menú contextual. Es necesario haber creado un espacio vacío previamente si no se quiere sobreescribir encima otros casos / variables. No pega el nombre del caso / variable.
- En el editor **no podemos deshacer** ni rehacer ninguna acción (si se borra cualquier información, por ejemplo, no se puede recuperar).
- Tampoco podemos hacer búsquedas.

Una vez introducidos los datos, o a medida que los vamos grabando para no perder el trabajo realizado, debemos guardarlos y convertirlos en un fichero del sistema R, por ejemplo con el nombre Encuesta.rda[15]. Para **guardar un archivo de datos**: a través del menú File / Save Data, haciendo clic sobre el botón o con las teclas <CTRL> <S>. Al guardar los datos nos aparecerá siempre la carpeta de trabajo definida por defecto (Mis documentos) o bien la que hayamos definido a través del menú File / Set Working Directory (<CTRL>+<D>). Es importante recordar que el directorio del archivo no puede tener acentos, ni tampoco el archivo de datos.

Los datos identificados de una matriz se corresponden con casos (filas) y variables (columnas). Éstas últimas se tratan como objetos del *workspace* del sistema R. Los objetos se pueden visualizar a través del menú: Packages & Data / Object Browser o clicando <CTRL><B> desde la consola. Esta opción permite también visualizar y editar las variables o incluso listar los datos con Print, pedir estadísticos de resumen a través de Summary o realizar gráficos con Plot. Lo podemos hacer del conjunto de las variables de la matriz o una a una.



En el caso de pedir un *summary* de toda la matriz de datos Encuesta se obtiene este resultado en la consola[16]:

[15] Esta matriz de datos se encuentra en la página web del capítulo.
[16] Se corresponde con el comando **`summary`** que vimos en el capítulo anterior.

```
> summary(Encuesta)
       ID              P1            P2                P3_1              P3_2
 Min.   : 1.00   Min.   :28.00   Varón:5   Sin            :0   Sin           :1
 1st Qu.: 3.25   1st Qu.:30.75   Mujer:5   EGB            :2   EGB           :3
 Median : 5.50   Median :35.00             FP1            :1   FP1           :1
 Mean   : 5.50   Mean   :34.90             Bachillerato   :3   Bachillerato  :3
 3rd Qu.: 7.75   3rd Qu.:39.25             FP2            :1   FP2           :2
 Max.   :10.00   Max.   :43.00             Universitarios:3   Universitarios:0

            P3_3         P4          P5                P6_1            P6_2            P6_3
 Sin           :2   Trabaja:5   Min.   :10.00   CDesacuerdo:0   CDesacuerdo:0   CDesacuerdo:1
 EGB           :3   No     :5   1st Qu.:16.25   Desacuerdo :3   Desacuerdo :1   Desacuerdo :2
 FP1           :1               Median :22.50   NiAniD     :2   NiAniD     :2   NiAniD     :3
 Bachillerato  :1               Mean   :23.33   Acuerdo    :3   Acuerdo    :4   Acuerdo    :2
 FP2           :3               3rd Qu.:28.75   CAcuerdo   :2   CAcuerdo   :3   CAcuerdo   :0
 Universitarios:0               Max.   :40.00                                    NA's       :2
                                NA's   :4
         P6_4            P6_5            P6_6            P6_7          P7
 CDesacuerdo:0   CDesacuerdo:1   CDesacuerdo:0   CDesacuerdo:0   Min.   :3.00
 Desacuerdo :3   Desacuerdo :0   Desacuerdo :1   Desacuerdo :1   1st Qu.:4.25
 NiAniD     :2   NiAniD     :1   NiAniD     :3   NiAniD     :3   Median :5.50
 Acuerdo    :4   Acuerdo    :3   Acuerdo    :2   Acuerdo    :1   Mean   :5.50
 CAcuerdo   :1   CAcuerdo   :5   CAcuerdo   :4   CAcuerdo   :5   3rd Qu.:6.00
                                                                 Max.   :9.00
```

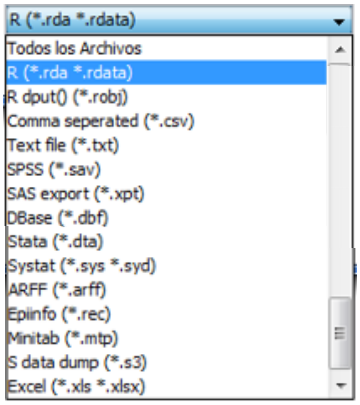
Una vez identificados los datos, un modo de comprobar la corrección del trabajo realizado es pedir las tablas de frecuencias a través del menú Analysis / Frequencies. Seleccionamos las variables y las pasamos en el recuadro de *Run Fequencies On* pulsando sobre el icono  . Finalmente ejecutamos el procedimiento de sacar las frecuencias pulsando sobre *OK*.

► Ejercicio 3. Propuesto
A partir de la matriz de datos creada Encuesta.rda obtener las tablas de frecuencias de las distintas variables y comprobar la correcta identificación de los datos.

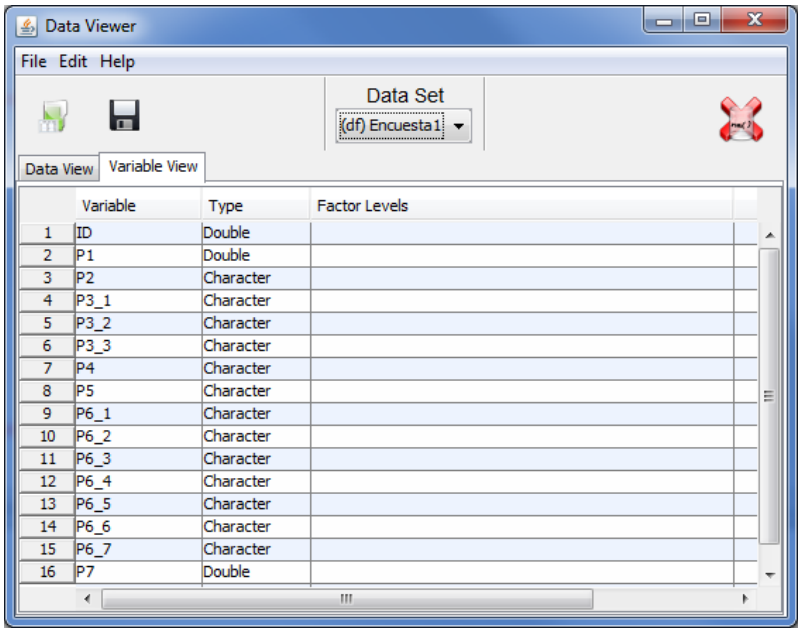
► Ejercicio 4. Propuesto
Con la matriz de datos CIS3041.rda obtener el diccionario de los datos y el libro de códigos para las variables: CCAA, TAMUNI, P3, P901, P1001, P1101, P1301, P15, P1601, P1701, P18, P2013, P23, P25, P28, P29, P31, P32, P46, VOTOSIM, RECUERDO, ESTUDIOS, OCUMAR11, CONDICION y ESTATUS, que permiten reconocer los principales tipos de variables y preguntas del Barómetro del CIS.
También se pueden pedir las tablas de frecuencias de todas ellas.

1.2.2. *Importación y exportación de datos en* R

Si disponemos de datos ya creados por otro software con un formato definido (SPSS, SAS, Excel,…) o bien sin formato, de texto plano (DAT, TXT), se puede importar fácilmente desde R. A través del menú File / Open Data de Deducer o con las teclas <CTRL>+<L>, o el botón  del *Data Viewer*, accedemos a un cuadro de diálogo que nos permite abrir un fichero eligiendo entre una diversidad de formatos:



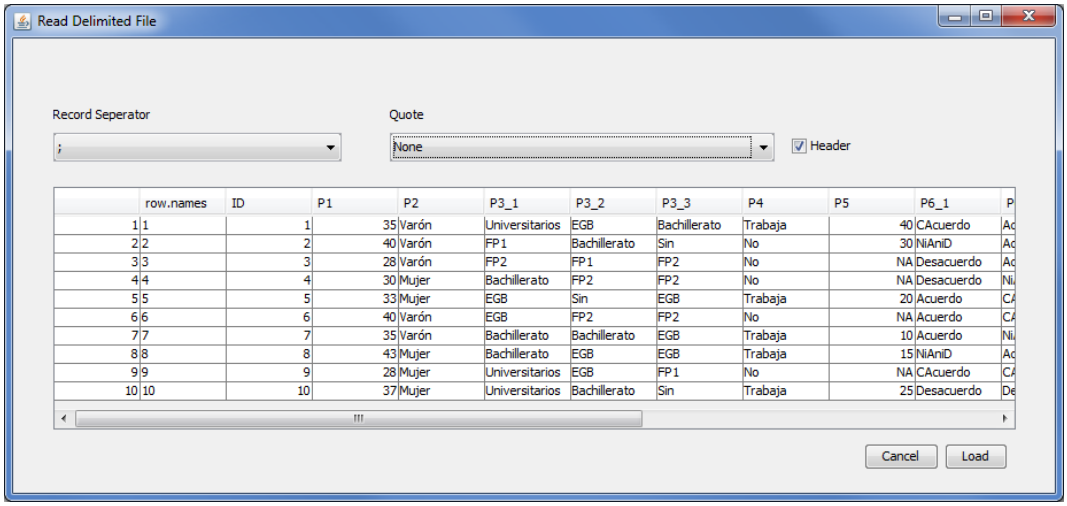
En la página web de este capítulo se encuentran los archivos Encuesta.xlsx, Encuesta.csv, Encuesta.sav y Encuesta.txt, si los abrimos desde Deducer veremos cómo realiza la importación. En el caso del archivo en formato Excel nos pedirá qué hoja de cálculo importar y a continuación creará una nueva matriz de datos con el nombre Encuesta1[17]. Se puede observar cómo ha asignado el nombre de las variables pues la primera línea de la hoja de Excel contiene el nombre y considera como variable de tipo *character* a los datos que están codificados textualmente. Cuando las convirtamos en variables tipo factor se generaran automáticamente los niveles o valores categóricos.



	Variable	Type	Factor Levels
1	ID	Double	
2	P1	Double	
3	P2	Character	
4	P3_1	Character	
5	P3_2	Character	
6	P3_3	Character	
7	P4	Character	
8	P5	Character	
9	P6_1	Character	
10	P6_2	Character	
11	P6_3	Character	
12	P6_4	Character	
13	P6_5	Character	
14	P6_6	Character	
15	P6_7	Character	
16	P7	Double	

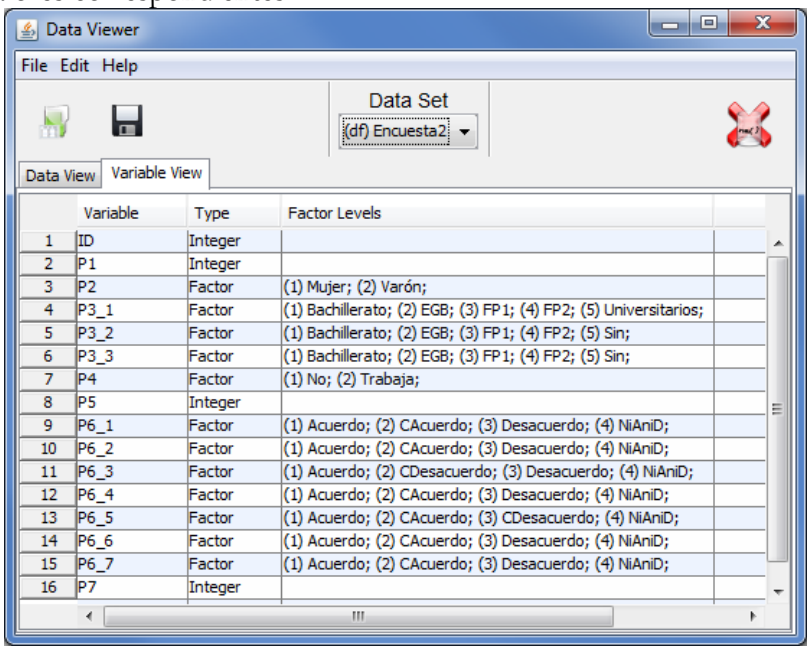
Podemos realizar en segundo lugar una importación de un archivo csv, es decir, un formato donde los datos están separados por una coma. Al abrir el fichero Encuesta.csv aparece este cuadro de diálogo de importación:

[17] Si estamos en un espacio de trabajo con la matriz Encuesta que hemos identificado.



	row.names	ID	P1	P2	P3_1	P3_2	P3_3	P4	P5	P6_1	P
1	1	1	35	Varón	Universitarios	EGB	Bachillerato	Trabaja	40	CAcuerdo	Ac
2	2	2	40	Varón	FP1	Bachillerato	Sin	No	30	NiAniD	Ac
3	3	3	28	Varón	FP2	FP1	FP2	No	NA	Desacuerdo	Ac
4	4	4	30	Mujer	Bachillerato	FP2	FP2	No	NA	Desacuerdo	Ni
5	5	5	33	Mujer	EGB	Sin	EGB	Trabaja	20	Acuerdo	CA
6	6	6	40	Varón	EGB	FP2	FP2	No	NA	Acuerdo	CA
7	7	7	35	Varón	Bachillerato	Bachillerato	EGB	Trabaja	10	Acuerdo	Ni
8	8	8	43	Mujer	Bachillerato	EGB	EGB	Trabaja	15	NiAniD	Ac
9	9	9	28	Mujer	Universitarios	EGB	FP1	No	NA	CAcuerdo	CA
10	10	10	37	Mujer	Universitarios	Bachillerato	Sin	Trabaja	25	Desacuerdo	De

Al cargarlo en R se genera la matriz de datos Encuesta2[18] con los datos, los nombres de las variables y las que son cualitativas ya se incorporan como variables tipo factor con sus valores correspondientes.

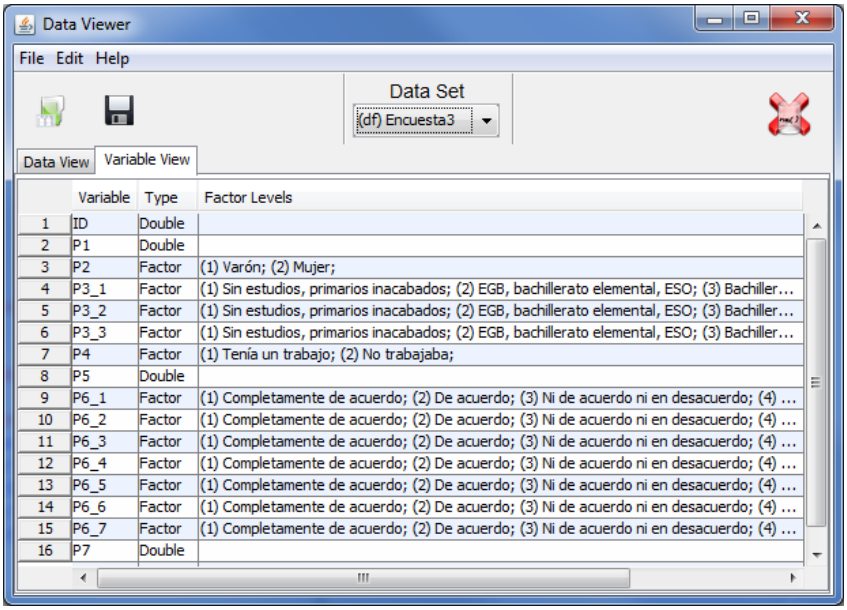


	Variable	Type	Factor Levels
1	ID	Integer	
2	P1	Integer	
3	P2	Factor	(1) Mujer; (2) Varón;
4	P3_1	Factor	(1) Bachillerato; (2) EGB; (3) FP1; (4) FP2; (5) Universitarios;
5	P3_2	Factor	(1) Bachillerato; (2) EGB; (3) FP1; (4) FP2; (5) Sin;
6	P3_3	Factor	(1) Bachillerato; (2) EGB; (3) FP1; (4) FP2; (5) Sin;
7	P4	Factor	(1) No; (2) Trabaja;
8	P5	Integer	
9	P6_1	Factor	(1) Acuerdo; (2) CAcuerdo; (3) Desacuerdo; (4) NiAniD;
10	P6_2	Factor	(1) Acuerdo; (2) CAcuerdo; (3) Desacuerdo; (4) NiAniD;
11	P6_3	Factor	(1) Acuerdo; (2) CDesacuerdo; (3) Desacuerdo; (4) NiAniD;
12	P6_4	Factor	(1) Acuerdo; (2) CAcuerdo; (3) Desacuerdo; (4) NiAniD;
13	P6_5	Factor	(1) Acuerdo; (2) CAcuerdo; (3) CDesacuerdo; (4) NiAniD;
14	P6_6	Factor	(1) Acuerdo; (2) CAcuerdo; (3) Desacuerdo; (4) NiAniD;
15	P6_7	Factor	(1) Acuerdo; (2) CAcuerdo; (3) Desacuerdo; (4) NiAniD;
16	P7	Integer	

Si importamos el archivo de SPSS Encuesta.sav, que difiere en la forma de haber etiquetado los valores de las variables cualitativas, vemos cómo se genera la matriz Encuesta3. En este caso se importan, como en el caso anterior, los nombres de las variables y las cualitativas como tipo factor con sus valores[19].

[18] Será así si estamos en un espacio de trabajo con la matriz Encuesta que hemos identificado al inicio y además hemos importado de Excel el archivo Encuesta.xlsx que pasó a denominarse Encuesta1.

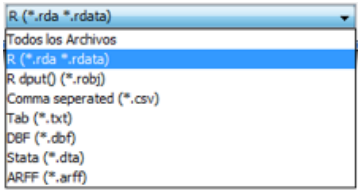
[19] La importación de variables de tipo fecha de SPSS genera problemas, por ello es mejor convertirlo a formato Excel e importarlo desde allí.



	Variable	Type	Factor Levels
1	ID	Double	
2	P1	Double	
3	P2	Factor	(1) Varón; (2) Mujer;
4	P3_1	Factor	(1) Sin estudios, primarios inacabados; (2) EGB, bachillerato elemental, ESO; (3) Bachiller…
5	P3_2	Factor	(1) Sin estudios, primarios inacabados; (2) EGB, bachillerato elemental, ESO; (3) Bachiller…
6	P3_3	Factor	(1) Sin estudios, primarios inacabados; (2) EGB, bachillerato elemental, ESO; (3) Bachiller…
7	P4	Factor	(1) Tenía un trabajo; (2) No trabajaba;
8	P5	Double	
9	P6_1	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
10	P6_2	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
11	P6_3	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
12	P6_4	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
13	P6_5	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
14	P6_6	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
15	P6_7	Factor	(1) Completamente de acuerdo; (2) De acuerdo; (3) Ni de acuerdo ni en desacuerdo; (4) …
16	P7	Double	

Finalmente podemos importar un archivo de texto plano como Encuesta.txt donde los datos están separados por tabulaciones. Los resultados son similares a los de la matriz importada Encuesta2.

También podemos guardar (exportar) nuestros datos en diferentes formatos. En este caso las opciones disponibles de formatos son menos pero suficientes para llevarlos a cualquier otra aplicación:

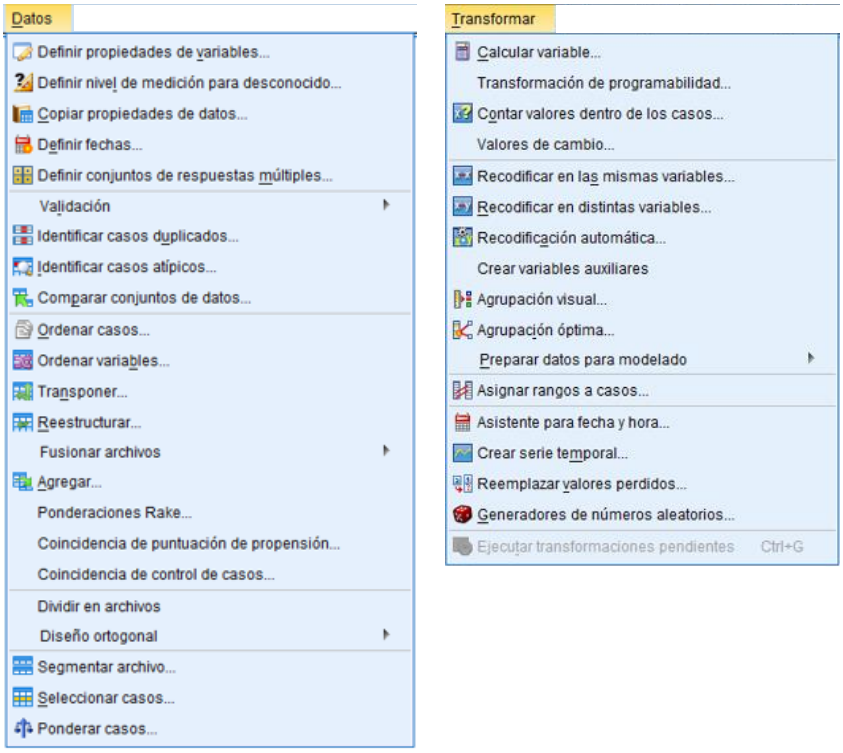


2. Transformación de los datos

La tarea de transformación de los datos está destinada a adaptar los datos a las necesidades del análisis donde se requiere modificarlos, para realizar correcciones y cambios en la información existente inicialmente, ya sea en relación a las variables de un archivo de datos o en relación al tratamiento de varios de ellos, o para generar nuevas variables basadas en las existentes: agrupaciones, tipologías, índices, etc. Como en el apartado anterior presentaremos en dos subapartados distintos los procedimientos de transformación para SPSS y R.

2.1. Transformación de los datos con SPSS

Comentaremos los distintos procedimientos que se presentan en los menús de SPSS Datos, destinado al tratamiento de ficheros, ya sea en su interior ya sea para combinarlo con otros, y Transformar, destinado a la transformación de las variables y a la creación de otras nuevas.



2.1.1. *Tratamiento de ficheros con SPSS*

Distinguiremos dos tipos de procedimientos de gestión y transformación de archivos, los destinados al tratamiento de datos en el interior de un fichero y al tratamiento de datos entre ficheros que se relacionan. Los comandos de SPSS que comentaremos son los de la Tabla III.2.2.

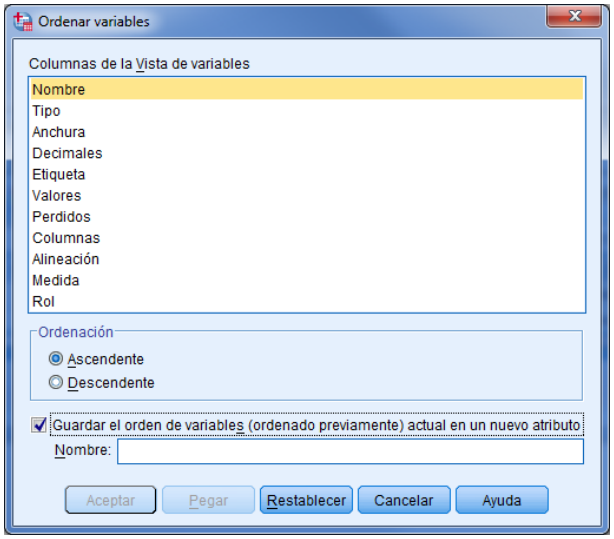
Tabla III.2.2 Procedimientos de tratamiento de ficheros

Menú Datos	Comandos de SPSS
Procedimientos de tratamiento de datos en el interior de un fichero	
Ordenar variables	**SORT VARIABLES**
Ordenar casos	**SORT CASES**
Seleccionar casos	**FILTER, SELECT IF, SAMPLE**
Segmentar archivo	**SPLIT FILE**
Ponderar casos	**WEIGHT**
Agregar	**AGGREGATE**
Transponer	**FLIP**
Reestructurar	**CASESTOVARS, VARSTOCASES**
Procedimientos de tratamiento de datos entre ficheros que se relacionan	
Dividir en archivos	**SPSSINC SPLIT DATASET**
Fusionar archivos	**MATCH FILES, ADD FILES**

2.1.1.1. *Tratamiento de datos en el interior de un fichero*

Ordenar variables

El comando **SORT VARIABLES** (menú Datos / Ordenar variables) puede ordenar las variables de la matriz en función de los valores de cualquiera de los atributos de variable del diccionario de los datos, de forma ascendente o descendente:

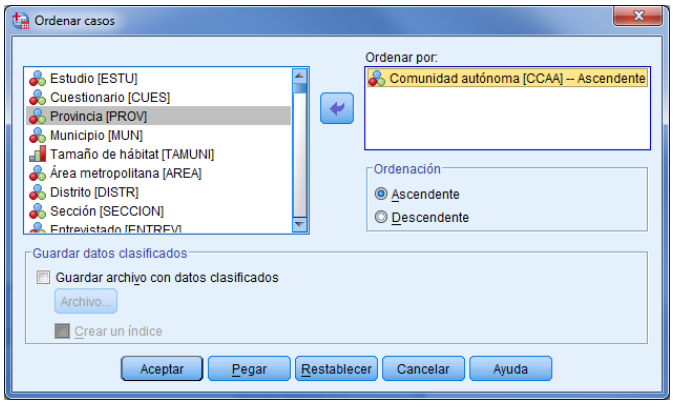


Es conveniente guardar el orden de las variables previo pues suele ser un criterio de ordenación que no se corresponde con ninguno preestablecido y podría ser difícil restaurarlo.

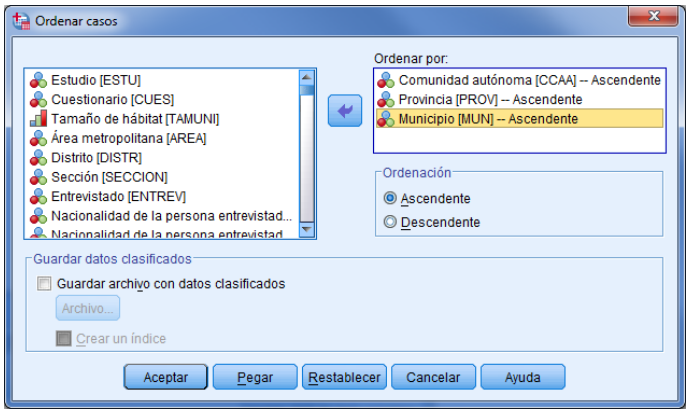
Ordenar casos

El comando **SORT CASES** (menú Datos / Ordenar casos) permite la reordenación de los casos del fichero activo según los valores especificados en una o más variables (hasta 10), numéricas o alfanuméricas (cadena, para éstas el orden es el alfabético). Los casos pueden ser reordenados en orden ascendente, por defecto, o descendente.

Con la matriz de datos CIS3041.sav vemos que los casos están inicialmente ordenados según el número del cuestionario (variable CUES). Como ejercicio podemos ordenar el archivo según el lugar de la entrevista. Un primer criterio sería por ejemplo ordenar el archivo según la Comunidad Autónoma (variable CCAA) en orden ascendente:



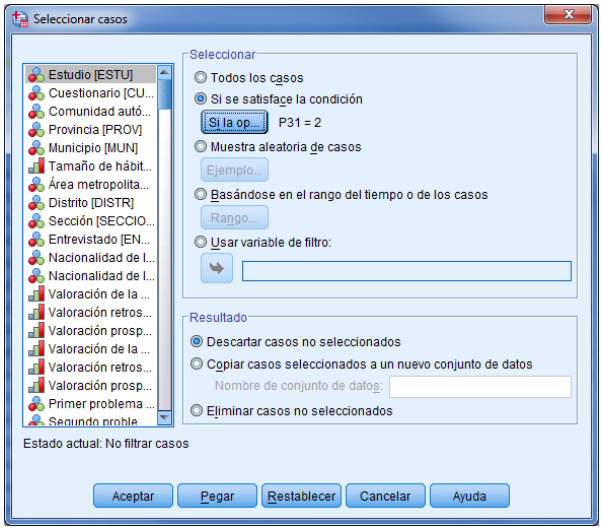
Obsérvense los cambios en el archivo de datos. Si queremos precisar más podemos poner además de la variable CCAA, la variable de la provincia (PROV) y del municipio (MUN), todas en orden ascendente. Las introduciremos por este orden:



Existe una opción para guardar en un archivo diferente los casos reordenados, con la posibilidad de crear un índice. La ordenación de un archivo de pequeñas dimensiones es instantánea pero con archivos de millones de registros puede tardar minutos, en este sentido es muy útil tener la base de datos ordenada según un criterio si se utiliza de forma habitual. Veremos también que la ordenación de un archivo es un paso previo necesario en diversos procedimientos de tratamiento de datos.

Seleccionar casos

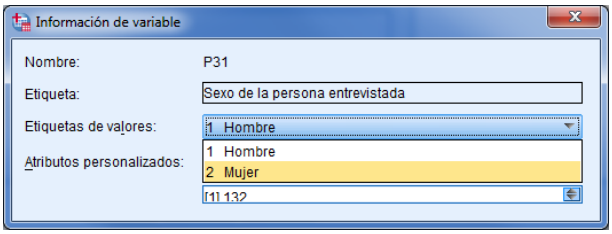
A menudo, cuando trabajamos con una base de datos nos interesa obtener información sobre los individuos que satisfacen determinadas condiciones. Nos puede interesar, por ejemplo, estudiar diversas variables pero sólo para los individuos con determinadas características: de sexo femenino, los que piensan votar, los que tienen un bajo nivel de ingresos, etc. El SPSS nos permite seleccionar los individuos que satisfacen una determinada condición de forma que, a partir de ese momento y mientras no deshacemos la selección, todos los procedimientos que aplicamos harán referencia sólo a los individuos seleccionados. Esta es la opción por defecto cuando elegimos Si se satisface la condición (opción Descartar casos no seleccionados) en el cuadro de diálogo de Datos / Seleccionar casos:



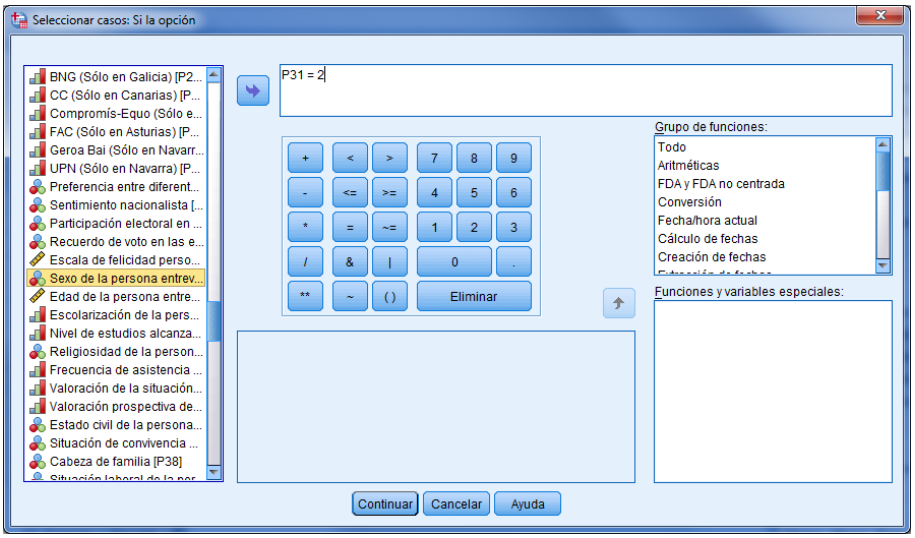
Esta operación se corresponde con el comando **FILTER**. Además de este procedimiento se posibilita la extracción de una muestra aleatoria de casos (comando **SAMPLE**), la selección a partir de un rango de casos (comando **USE**), y utilizar variables filtro. En cualquiera de estos casos podemos optar por:

- Descartar casos no seleccionados: la selección implica que los datos son filtrados, es decir, los casos filtrados permanecen en el archivo pero se excluyen del análisis y se pueden recuperar. Habitualmente se trabaja de esta manera.
- Copiar los casos seleccionados a un nuevo archivo de datos.
- Eliminar casos no seleccionados: se eliminan los casos no seleccionados del archivo activo (el de la memoria temporal del sistema). El archivo original se mantiene en el disco, pero si después de hacer la selección guardamos el archivo con el mismo nombre entonces perderemos definitivamente los casos no seleccionados.

Como ejercicio podemos seleccionar los casos de las personas entrevistadas que son mujeres. Elegimos Si se satisface la condición y pulsamos sobre el icono de Si la op.... En el nuevo cuadro de diálogo construiremos la condición[20]. Seleccionamos variable del sexo (la P31) y la pasamos a la derecha. Para seleccionar a las mujeres escribiremos con el teclado o con los botones del cuadro de diálogo: = 2. El valor 2 corresponde a las mujeres. En el caso de que no recordáramos el código, una forma inmediata de consultarlo es darle al botón derecho del ratón y clicar sobre Información de variable:

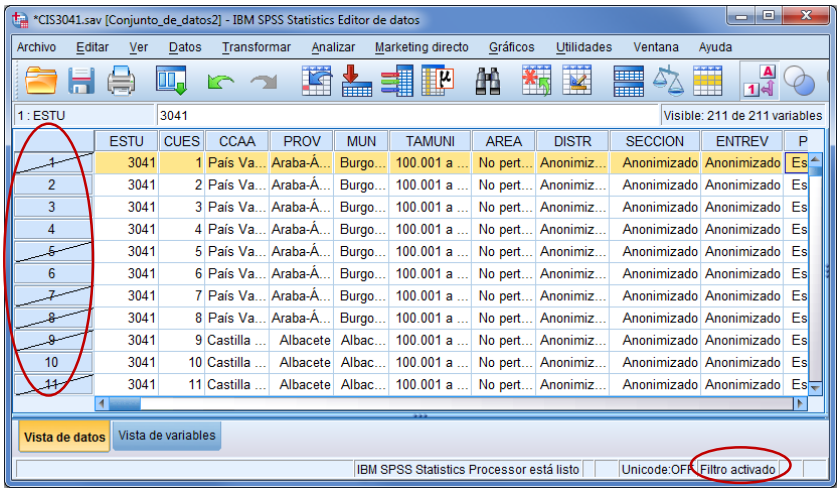


Construida la condición:



[20] Este cuadro de diálogo los veremos también en el procedimiento Calcular para transformar los datos. Para establecer una condición es necesario manejar expresiones de transformación que comentaremos en el apartado siguiente.

Clicaremos sobre Continuar y sobre Aceptar en el siguiente cuadro de diálogo para que realice la acción, asegurándonos de que esté activada la opción Descartar. Si observamos ahora la base de datos, veremos que aparecen algunos casos "tachados" en el margen izquierdo de numeración del caso: son los casos que no han sido seleccionados, es decir, los individuos hombres.



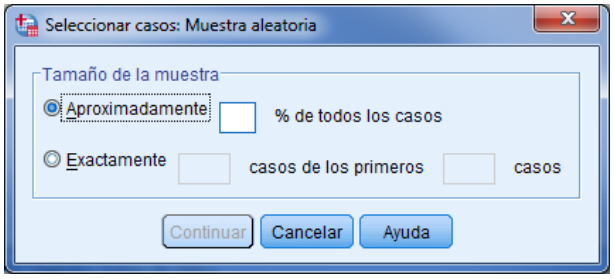
Obsérvese igualmente que se ha creado automáticamente una nueva variable de filtro, la última de la matriz de datos, llamada filter_$ que toma los valores 0 y 1 con etiquetas Not selected y Selected, respectivamente, según si el individuo ha sido seleccionado o no. Obsérvese también que en la parte inferior derecha de la ventana del SPSS aparece una etiqueta con la inscripción Filtro activado. Nos recuerda que el archivo de datos con el que trabajamos ha sido filtrado, es decir, nos recuerda que no estamos trabajando con todos los datos sino sólo con las que satisfacen una determinada característica. También nos ha aparecido en el archivo de resultados la anotación de los comandos de sintaxis indicando que se han filtrado los casos.

Si ahora calculamos, por ejemplo, la tabla de frecuencias de una variable cualquiera, la información obtenida se referirá sólo a las mujeres de nuestra base de datos. Es muy importante que, una vez hayamos realizado el estudio que queríamos hacer con sólo una parte de los individuos, nos acordemos de deshacer la selección para volver a trabajar con el archivo completo. Si no lo hiciéramos estaríamos obteniendo informaciones erróneas. Para ello volveríamos al menú de la selección y marcaríamos la opción Todos los casos.

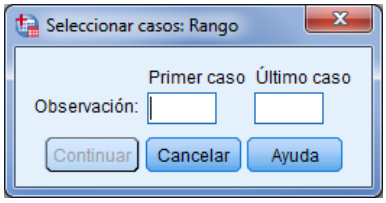
Si la ejecución del procedimiento de selección que acabamos de realizar se hubiera hecho con la opción Eliminar casos no seleccionados, entonces estaríamos ejecutando otro comando de SPSS, el que corresponde a **SELECT IF**[21].

Si quisiéramos extraer una muestra aleatoria de casos especificaríamos en su cuadro de diálogo un % aproximado o un número de casos dado:

[21] Cuando se elaboran programas de sintaxis se puede escribir el comando **SELECT IF** precedido de **TEMPORARY**, así se consigue aplicar una selección temporal que afecta solamente al siguiente comando de procedimiento, después se vuelven a considerar todos los casos.



En el caso de definir un rango de casos el cuadro de diálogo sería el siguiente:



Todos estos procedimientos se corresponden con comandos de transformación, es decir, comandos que no realizan la tarea (no acceden a la lectura de los datos) si no encuentran un comando que fuerce la lectura de los datos (cualquier procedimiento de análisis por ejemplo). Cuando se ejecutan por el menú estos comandos su acción se realiza inmediatamente porque se adjunta en la ejecución un comando adicional: **EXECUTE**, como puede observarse en el archivo de resultados, destinado a obligar a la lectura de los datos y realizar todas las acciones de transformación que hubieran hasta ese momento[22].

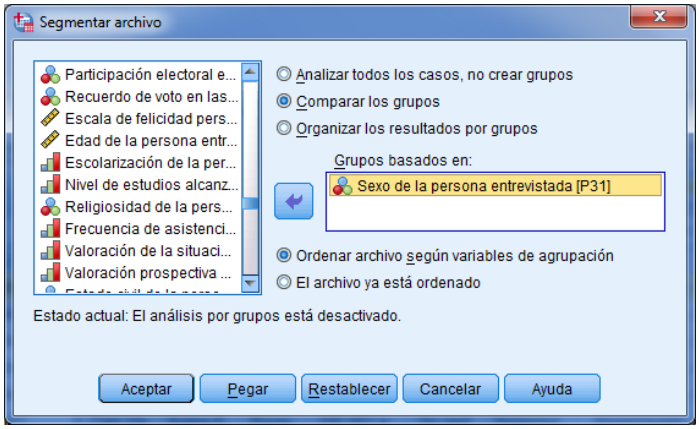
Segmentar archivo

Otra necesidad habitual en el tratamiento de los datos de un fichero es segmentarlo, es decir, dividirlo en grupos de individuos según los valores de una o más variables de agrupación para realizar un mismo tipo de análisis que se repetirá dentro de cada grupo. Para poder realizar a la segmentación correctamente será necesario ordenar previamente el archivo. El SPSS nos ofrece dos formas diferentes de segmentar el archivo:

- Comparar los grupos: los grupos se presentan juntos para poder compararlos en una sola tabla o con gráficos individuales que se presentan juntos.
- Organizar los resultados por grupos: los resultados de cada procedimiento se muestran por separado para cada grupo.

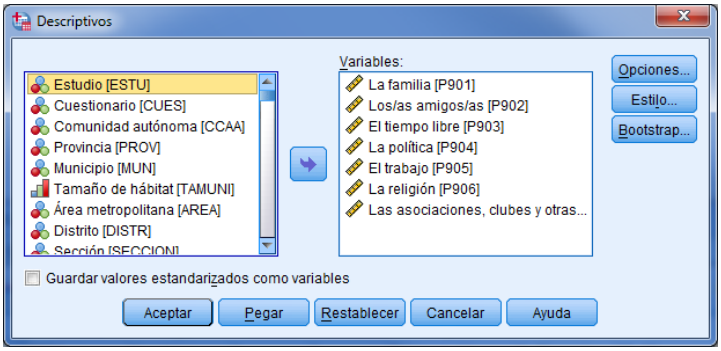
El comando de segmentación es **SPLIT FILE** (menú Datos / Segmentar archivo). El cuadro de diálogo inicial es:

[22] Ver en capítulo anterior el apartado sobre el lenguaje de comandos de SPSS donde se explica el concepto de estados del programa.



En él podemos ver que se ha introducido la variable de segmentación sexo (P31) y aparece marcada la opción por defecto Comparar grupos. Si nuestro archivo de datos no está ordenado por la variable de segmentación marcaremos que lo ordene previamente pues es una condición necesaria agrupar primero los individuos. Ejecutaremos esta transformación del archivo y veremos que en la parte inferior derecha de la ventana del SPSS aparece una etiqueta con la inscripción Dividir por.

A partir de ese momento todo ejercicio de análisis que ejecutemos se realizará para cada grupo. Por ejemplo podemos solicitar, los descriptivos de las variables a través del menú Analizar / Estadísticos descriptivos / Descriptivos de las variables P901 a P907:



El resultado es el siguiente:

Estadísticos descriptivos

P31 Sexo de la persona entrevistada		N	Mínimo	Máximo	Media	Desviación estándar
1 Hombre	P901 La familia	1208	0	10	9,56	1,134
	P902 Los/as amigos/as	1203	0	10	8,16	1,833
	P903 El tiempo libre	1193	0	10	7,76	1,978
	P904 La política	1199	0	10	4,03	3,262
	P905 El trabajo	1193	0	10	8,78	2,096
	P906 La religión	1193	0	10	3,31	3,198
	P907 Las asociaciones, clubes y otras actividades asociativas	1145	0	10	5,11	2,760
	N válido (por lista)	1110				
2 Mujer	P901 La familia	1266	0	10	9,76	,832
	P902 Los/as amigos/as	1258	0	10	8,18	1,878
	P903 El tiempo libre	1244	0	10	7,86	1,902
	P904 La política	1241	0	10	3,79	3,313
	P905 El trabajo	1243	0	10	8,92	1,935
	P906 La religión	1251	0	10	4,53	3,507
	P907 Las asociaciones, clubes y otras actividades asociativas	1168	0	10	5,09	2,859
	N válido (por lista)	1126				

Una única tabla con el análisis realizado para hombres y mujeres. Si volvemos a ejecutar el procedimiento con la opción Organizar los resultados por grupos, obtendremos la misma información pero en tablas separadas.

Sexo de la persona entrevistada = Hombre

Estadísticos descriptivos[a]

	N	Mínimo	Máximo	Media	Desviación estándar
P901 La familia	1208	0	10	9,56	1,134
P902 Los/as amigos/as	1203	0	10	8,16	1,833
P903 El tiempo libre	1193	0	10	7,76	1,978
P904 La política	1199	0	10	4,03	3,262
P905 El trabajo	1193	0	10	8,78	2,096
P906 La religión	1193	0	10	3,31	3,198
P907 Las asociaciones, clubes y otras actividades asociativas	1145	0	10	5,11	2,760
N válido (por lista)	1110				

a. P31 Sexo de la persona entrevistada = 1 Hombre

Sexo de la persona entrevistada = Mujer

Estadísticos descriptivos[a]

	N	Mínimo	Máximo	Media	Desviación estándar
P901 La familia	1266	0	10	9,76	,832
P902 Los/as amigos/as	1258	0	10	8,18	1,878
P903 El tiempo libre	1244	0	10	7,86	1,902
P904 La política	1241	0	10	3,79	3,313
P905 El trabajo	1243	0	10	8,92	1,935
P906 La religión	1251	0	10	4,53	3,507
P907 Las asociaciones, clubes y otras actividades asociativas	1168	0	10	5,09	2,859
N válido (por lista)	1126				

a. P31 Sexo de la persona entrevistada = 2 Mujer

Esta opción tiene diversas aplicaciones, pero una de ellas podría ser la de elaborar el anexo estadístico con numerosas tablas y gráficos que queremos repetir, por ejemplo, para cada territorio del estudio por separado.

Aquí de nuevo es importante recordar que una vez hayamos realizado el análisis deseado es necesario deshacer la segmentación para volver a trabajar con el archivo completo, como una sola muestra. Para ello volvemos al menú y marcamos Analizar todos los casos.

Ponderar casos

La ponderación de los datos es otra de las necesidades recurrentes de un análisis cuantitativo de datos. Si se ponderan los casos lo que hacemos es cambiar el peso que tiene cada caso. Por defecto cada individuo vale una unidad y el recuento de cualquier característica, por ejemplo ser hombre, es la suma de tantos 1 como individuos tienen ese valor. Pero el valor del peso de cada individuo se puede cambiar, y ello significa cambiar una variable interna del sistema SPSS de nombre $weight. Esta variable interna siempre vale 1 para cada individuo hasta que la cambiamos con el comando de ponderación **WEIGHT** o por el menú Datos / Ponderar casos.

La necesidad de ponderar se puede presentar en diferentes situaciones. Comentaremos tres de ellas. Una primera situación muy habitual tiene que ver con la necesidad de

ponderar los datos de una muestra, ya sea por el propio diseño de construcción[23] o porque se tiene la necesidad de equilibrarla dado que se han podido constar ciertos desequilibrios o sesgos en la información recogida. Imaginemos por ejemplo que la proporción poblacional de varones y mujeres en un territorio fuera de 50 y 50 por ciento, obtenemos una muestra de esa población y nos sale 48 y 52. Nuestros resultados tendrán un sesgo en favor de los perfiles de las mujeres que apareen un 2% más de lo que corresponde. Para corregir este desvío y restituir el 50% de su población en términos muestrales es necesario introducir una ponderación de tal manera que convierta el peso de los hombre de 48 a 50 y el de las mujeres de 52 a 50.

Si nuestra muestra es de 1000 individuos eso implica que tenemos 480 varones y 520 mujeres, la ponderación se genera aplicando la fórmula siguiente:

$$w_i = \frac{peso\ teórico}{peso\ real}$$

En el caso de los varones ($i=1$) teóricamente deberían ser el 50%, es decir, 500 individuos, pero el peso real es de 480, quiere decir por tanto que debemos aumentar la importancia de los varones multiplicando cada individuo por un valor superior a 1, en concreto, 1,083.

$$w_{varones} = \frac{520}{480} = 1{,}083$$

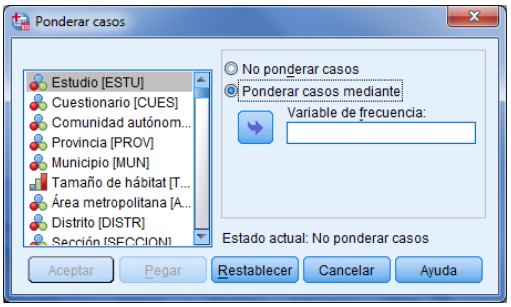
El mismo razonamiento en el caso de las mujeres genera un peso inferior a 1 de 0,923:

$$w_{mujeres} = \frac{480}{520} = 0{,}923$$

Si a cada varón lo multiplicamos por 1,083 en vez 1 y a cada mujer por 0,923 en vez de 1, en el recuento final tendremos 500 varones y 500 mujeres. Para hacerlo efectivo en el SPSS es necesario crear primero la variable de ponderación y después ponderar. Veremos en el próximo apartado cómo generar variables. Si lo hiciéramos por sintaxis sería por ejemplo así:

```
IF sexo=1 peso=1.083.
IF sexo=2 peso=0.923.
WEIGHT BY peso.
```

El comando de ponderación es muy sencillo, y su cuadro de diálogo es el siguiente, donde solo se trata de elegir la variable de ponderación:



[23] La ponderación a veces también se acompaña de la necesidad de elevar la muestra, es decir de expresar los individuos de la muestra en términos poblacionales por lo que se multiplica cada individuo por lo que valga en términos poblacionales. Así expresan también los datos por ejemplo de la Encuesta de Población Activa. Ponderar y elevar son dos pesos y dos ponderaciones que se pueden aplicar simultáneamente o por separado.

Un segundo ejercicio de ponderación lo haremos con datos cuyas unidades son agregadas. Es el caso de la matriz sobre el índice de desarrollo humano IDH2014.sav donde cada unidad es un país. Cuando trabajamos con este archivo, si no ponderamos los casos, todos los países tienen el mismo peso, independientemente de su población, superficie, etc. A veces nos interesará trabajar con el archivo de esta manera, pero en otros casos puede ser erróneo. Si queremos analizar, por ejemplo, cuál es el producto interior bruto per cápita mundial, no podemos dar el mismo peso a Andorra (0,08 millones) que a China (1.385,57millones). En este caso sería conveniente dar a cada país un peso diferente según su población, proporcional al número de persona que habitan en el país.

Empezaremos calculando la media de la variable GDPpercapita (*Gross Domestic Product per capita*) sin ponderar los casos. Obtenemos el siguiente resultado:

Estadísticos descriptivos

	N	Media
GDPpercapita GDP per capita	180	16496,9136
N válido (por lista)	180	

16.497$ es una media donde los individuos son países. A partir de la riqueza de cada país hemos calculado la media dando el mismo peso a todos los países. Por tanto no es un reflejo exacto del producto interior bruto per cápita mundial. Para calcularla debemos dar a cada país un peso proporcional a su población. Ponderamos a través del menú Datos / Ponderar casos / Ponderar casos mediante y escogemos la variable Population que nos da la población de cada país en millones. El nuevo cálculo de la media arroja este resultado:

Estadísticos descriptivos

	N	Media
GDPpercapita GDP per capita	6951	13552,3587
N válido (por lista)	6951	

Obsérvese que la media ahora ha bajado a 13.552$, antes teníamos 180 países y ahora el valor es de 6.951 personas (la población mundial en millones). Este resultado aproxima mucho mejor el PIB per cápita mundial al tener en cuenta los países más poblados que mayormente son menos ricos por lo que la media mundial baja.

Un vez realizado un análisis ponderando los casos debemos recordar deshacerla si no la necesitamos. En caso contrario obtendríamos información incorrecta. Para ello volvemos al menú: Datos / Ponderar casos / No ponderar los casos.

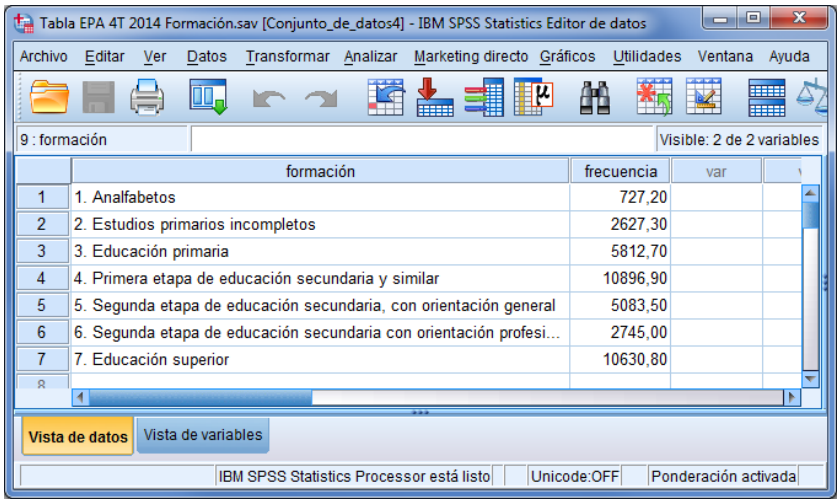
El comando de ponderación se puede utilizar también instrumentalmente para reproducir tablas de frecuencias de una o más variables. Por ejemplo, si entramos en la web del Instituto Nacional de Estadística y consultamos los datos de la Encuesta de Población Activa del 4º trimestre de 2014 podemos ver entre otros muchos datos que la distribución de la población según el nivel de estudios alcanzado es la siguiente:

Encuesta de Población Activa	
Población en viviendas familiares	
Población de 16 y más años y nivel de formación alcanzado	
Unidades: Miles Personas	
	Total
	2014T4
Total	38.523,4
Analfabetos	727,2
Estudios primarios incompletos	2.627,3
Educación primaria	5.812,7
Primera etapa de educación secundaria y similar	10.896,9
Segunda etapa de educación secundaria, con orientación general	5.083,5
Segunda etapa de educación secundaria con orientación profesional	2.745,0
Educación superior	10.630,8

Fuente: Instituto Nacional de Estadística, EPA 2014

Los datos de la encuesta están elevados a toda la población y hacen referencia a miles de personas. En total la población de 16 y más años es de 38.523.400 personas que se distribuyen según las 7 categorías del nivel de formación. Si queremos trabajar con estos datos, por ejemplo, para extraer una tabla de frecuencias relativas o elaborar un gráfico, en una ventana de datos en blanco podemos introducir dos variables: una con los diferentes niveles de estudios (variable formación) y otra con la frecuencia, la variable que actúa de peso (variable frecuencia), es decir, con el número de individuos de cada categoría, variable con la que ponderaremos los casos.

La ventana de datos de SPSS quedaría así:

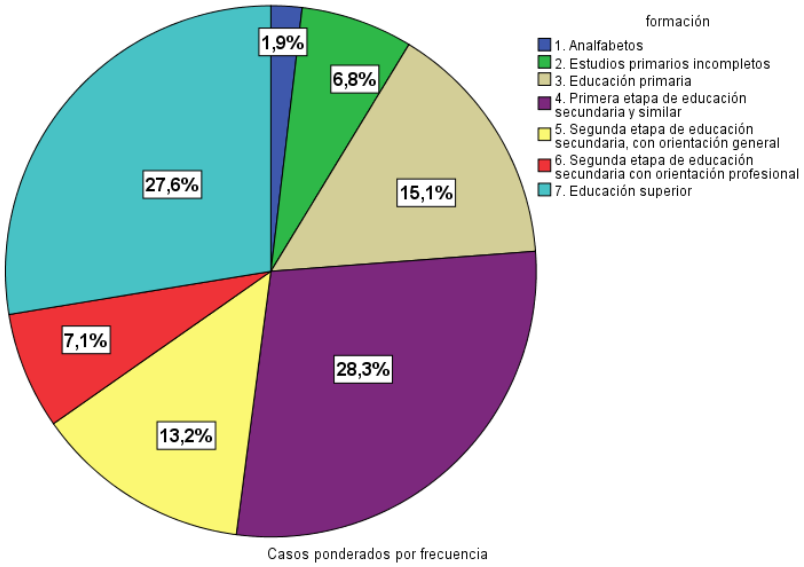


Una vez hecho esto se ponderan los casos según la variable frecuencia. En la parte inferior derecha de la ventana del SPSS aparecerá una etiqueta con la inscripción Ponderación activada. A partir de ese momento el número de casos que tenemos, 7, donde cada caso valía 1, tras la ponderación, pasa a valer el número de casos que indique la columna frecuencia, y en total los 38 millones y medio de la tabla original. Podemos ejecutar el procedimiento Frecuencias para la variable formación y obtenemos reproducida la tabla de la EPA:

formación

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1. Analfabetos	727	1,9	1,9	1,9
	2. Estudios primarios incompletos	2627	6,8	6,8	8,7
	3. Educación primaria	5813	15,1	15,1	23,8
	4. Primera etapa de educación secundaria y similar	10897	28,3	28,3	52,1
	5. Segunda etapa de educación secundaria, con orientación general	5084	13,2	13,2	65,3
	6. Segunda etapa de educación secundaria con orientación profesional	2745	7,1	7,1	72,4
	7. Educación superior	10631	27,6	27,6	100,0
	Total	38523	100,0	100,0	

Y un gráfico de sectores por ejemplo:

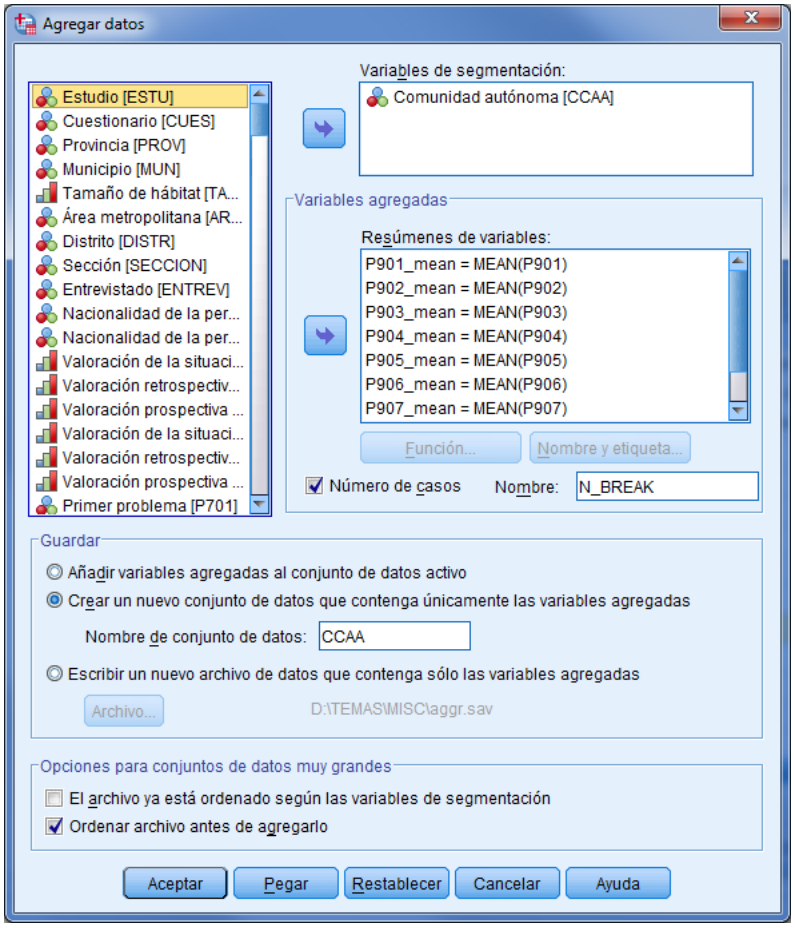


Agregar

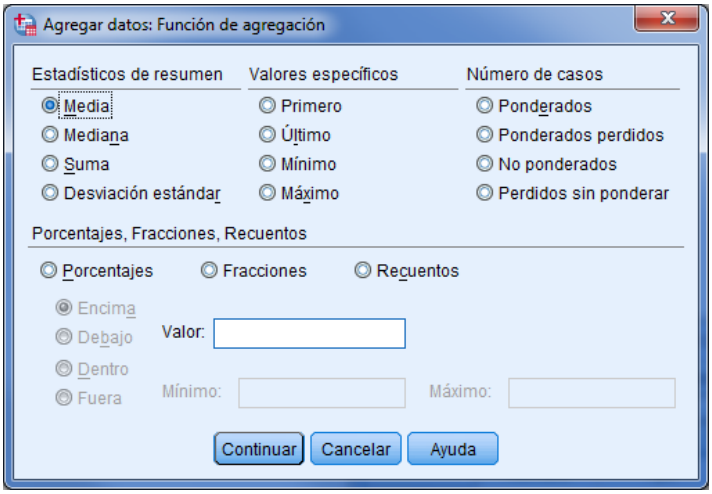
La agregación de casos tiene múltiples usos en el tratamiento de matrices de datos, también cuando se relacionan diversas bases. Es especialmente útil cuando disponemos de información en matrices distintas con diferentes niveles de agregación, como en el caso de tener información de individuos y de hogares en la *Encuesta de Población Activa*, o de tener múltiples registros de la vida laboral para un mismo individuo del que tenemos información sociodemográfica en otra base, como en la *Muestra Continua de Vidas Laborales de la Seguridad Social*.

Veremos un ejercicio sencillo de aplicación para ver cómo funciona el procedimiento. Se tratará de agregar los individuos entrevistados en la encuesta del CIS según su Comunidad Autónoma calculando una medida de resumen (la media) de las variables: P001 a P907 (Importancia de diversos aspecto de la vida social), P30 (Escala de felicidad personal) y P32 (Edad).

Se agrega con el comando **AGGREGATE** (menú Datos / Agregar). En el cuadro de diálogo debemos determinar en primer lugar la o las variables que actúan de segmentación, es decir, los grupos de agregación. En nuestro caso elegimos la Comunidad Autónoma, por tanto, tendremos 19 grupos.



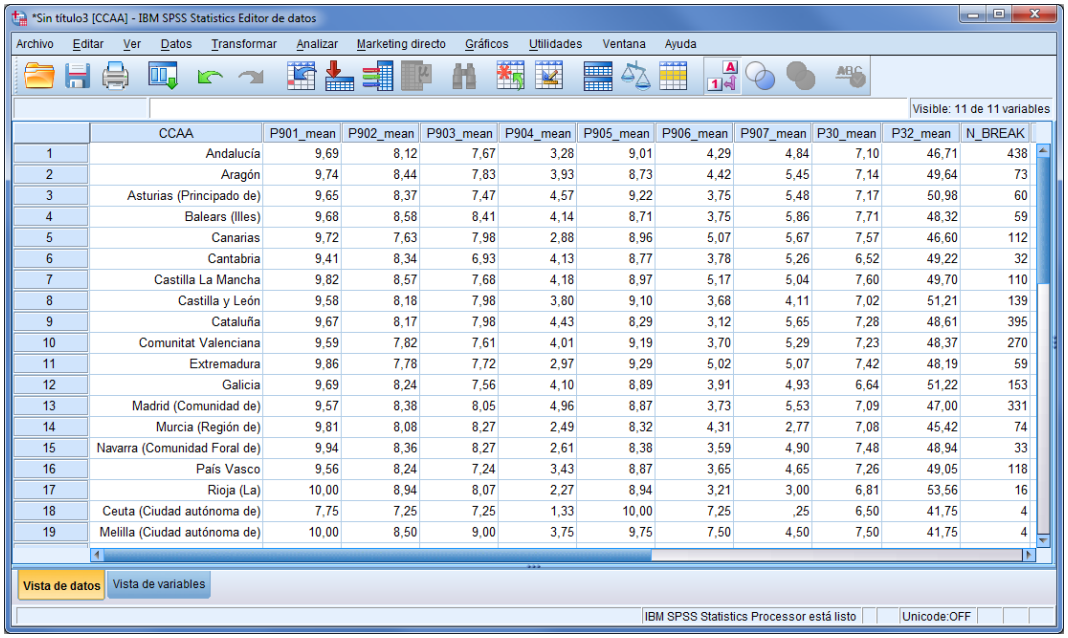
Dentro de cada grupo podemos calcular distintas medidas de resumen. Para ello elegimos primero las variables de interés y las pasamos al recuadro de Variables agregadas, automáticamente el sistema SPSS elige la media como medida, pero podemos cambiarla eligiendo una o varias variables y clicando a continuación sobre Función. Accederemos al cuadro de diálogo que permite elegir la función. En nuestro caso dejaremos el estadístico de la media. Cada nuevo cálculo genera una variable que se puede definir con un nombre específico y una etiqueta, sino SPSS nos ofrece el criterio Nombre-variable_estadístico. Un cálculo adicional permite añadir la variable con el número de casos de cada grupo, que por defecto tiene el nombre de N_BREAK.



Definidos los cálculos podemos optar por tres alternativas:

- Añadir variables agregadas al conjunto de datos activo. Las nuevas variables calculadas de grupo son un atributo de cada unidad de la base de datos original por lo que cada caso con los mismos valores de segmentación recibe los mismos valores para las nuevas variables agregadas.
- Crear un nuevo conjunto de datos que contenga únicamente las variables agregadas. Se crea un nuevo conjunto de datos en la sesión actual con las variables de agregación y agrega las unidades.
- Escribir un nuevo archivo de datos que contenga sólo las variables agregadas. Es el caso anterior pero guarda los datos agregados en un archivo de datos externo que hay que detallar.

En nuestro ejercicio elegimos la segunda opción y obtenemos una una matriz de datos que contiene las 19 líneas con cada Comunidad Autónoma y 10 variables nuevas que calculan la media de las variable P901 a P907, P30, P32 más N_NREAK.



	CCAA	P901_mean	P902_mean	P903_mean	P904_mean	P905_mean	P906_mean	P907_mean	P30_mean	P32_mean	N_BREAK
1	Andalucía	9,69	8,12	7,67	3,28	9,01	4,29	4,84	7,10	46,71	438
2	Aragón	9,74	8,44	7,83	3,93	8,73	4,42	5,45	7,14	49,64	73
3	Asturias (Principado de)	9,65	8,37	7,47	4,57	9,22	3,75	5,48	7,17	50,98	60
4	Balears (Illes)	9,68	8,58	8,41	4,14	8,71	3,75	5,86	7,71	48,32	59
5	Canarias	9,72	7,63	7,98	2,88	8,96	5,07	5,67	7,57	46,60	112
6	Cantabria	9,41	8,34	6,93	4,13	8,77	3,78	5,26	6,52	49,22	32
7	Castilla La Mancha	9,82	8,57	7,68	4,18	8,97	5,17	5,04	7,60	49,70	110
8	Castilla y León	9,58	8,18	7,98	3,80	9,10	3,68	4,11	7,02	51,21	139
9	Cataluña	9,67	8,17	7,98	4,43	8,29	3,12	5,65	7,28	48,61	395
10	Comunitat Valenciana	9,59	7,82	7,61	4,01	9,19	3,70	5,29	7,23	48,37	270
11	Extremadura	9,86	7,78	7,72	2,97	9,29	5,02	5,07	7,42	48,19	59
12	Galicia	9,69	8,24	7,56	4,10	8,89	3,91	4,93	6,64	51,22	153
13	Madrid (Comunidad de)	9,57	8,38	8,05	4,96	8,87	3,73	5,53	7,09	47,00	331
14	Murcia (Región de)	9,81	8,08	8,27	2,49	8,32	4,31	2,77	7,08	45,42	74
15	Navarra (Comunidad Foral de)	9,94	8,36	8,27	2,61	8,38	3,59	4,90	7,48	48,94	33
16	País Vasco	9,56	8,24	7,24	3,43	8,87	3,65	4,65	7,26	49,05	118
17	Rioja (La)	10,00	8,94	8,07	2,27	8,94	3,21	3,00	6,81	53,56	16
18	Ceuta (Ciudad autónoma de)	7,75	7,25	7,25	1,33	10,00	7,25	,25	6,50	41,75	4
19	Melilla (Ciudad autónoma de)	10,00	8,50	9,00	3,75	9,75	7,50	4,50	7,50	41,75	4

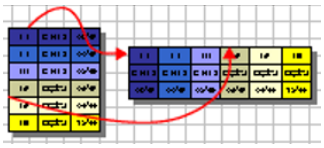
En este procedimiento también es necesario tener previamente los casos de la matriz original ordenados según las variables de segmentación.

Transponer

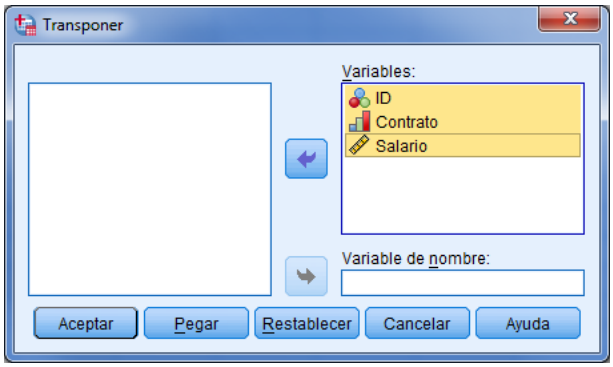
La transposición de una matriz implica convertir los casos (las filas) en variables, y las variables (las columnas) en casos. Al hacerlo se crea un nuevo archivo de datos y automáticamente los nombres de las variables.

Para ilustrar este comando, **FLIP** (menú Datos / Transponer), y los que vienen a continuación, trabajaremos con unas pequeñas matrices de datos que permitirán ver mejor cada una de las tareas. La matriz de datos X.sav contiene la situación laboral de 6 individuos asalariados en relación a 2 variables de sus condiciones de empleo: Contrato y Salario.

	ID	Contrato	Salario
1	1	Fijo	1200
2	2	Temporal	1000
3	3	Fijo	3000
4	4	Temporal	1000
5	5	Fijo	1200
6	6	Fijo	1500



En el menú pasamos todas las variables al recuadro de la derecha y ejecutamos:



El resultado obtenido es el siguiente:

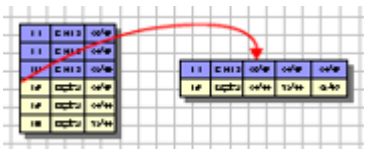
	CASE_LBL	var001	var002	var003	var004	var005	var006
1	ID	1,00	2,00	3,00	4,00	5,00	6,00
2	Contrato	1,00	2,00	1,00	2,00	1,00	1,00
3	Salario	1200,00	1000,00	3000,00	1000,00	1200,00	1500,00

Reestructurar

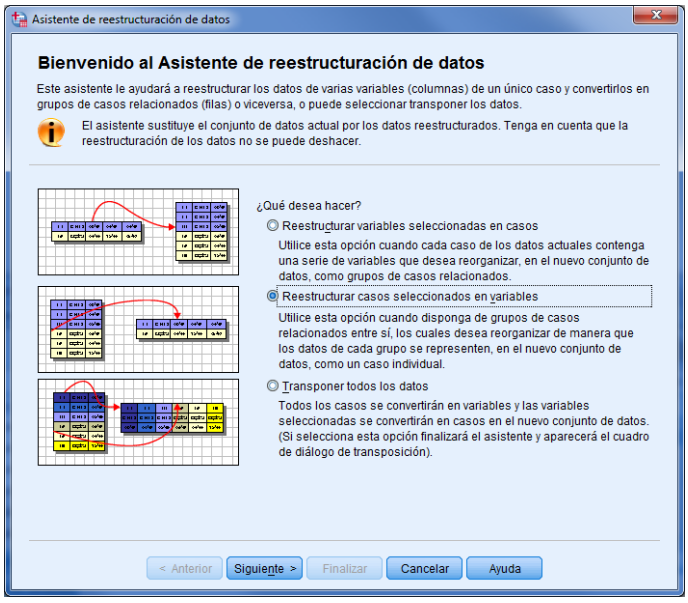
La estructura simple de una matriz de datos de casos por variables suele ser la habitual para el análisis de datos, no obstante, la estructura inicial de una base de datos puede ser compleja. Una estructura simple es el ejemplo de la matriz X.sav, de 6 individuos y 2 variables con las condiciones de empleo. Una estructura donde la información de una variable está en más de una columna o la información de un caso en más de una fila introduce una complejidad de organización de la información y la necesidad de reestructurar el archivo para pasar los casos a variables o las variables a casos.

Por ejemplo, si tenemos una matriz con 3 individuos y las condiciones de empleo se refieren a dos momentos en el tiempo: empleo inicial y empleo actual, la información puede estar dispuesta por filas donde cada individuo tiene doble información de sus condiciones de empleo, la inicial y la actual. La matriz de datos casestovars.sav tiene esta información:

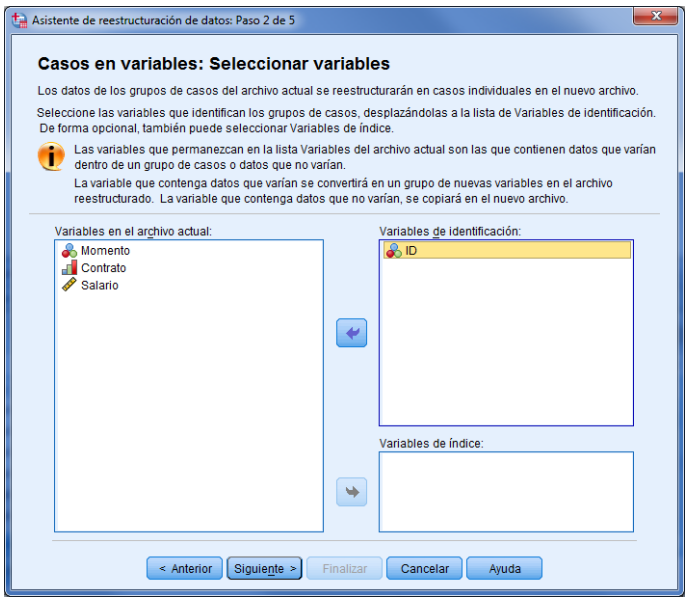
	ID	Momento	Contrato	Salario
1	1	Inicial	Temporal	1000
2	1	Actual	Fijo	1200
3	2	Inicial	Fijo	1500
4	2	Actual	Temporal	1000
5	3	Inicial	Fijo	2000
6	3	Actual	Fijo	3000



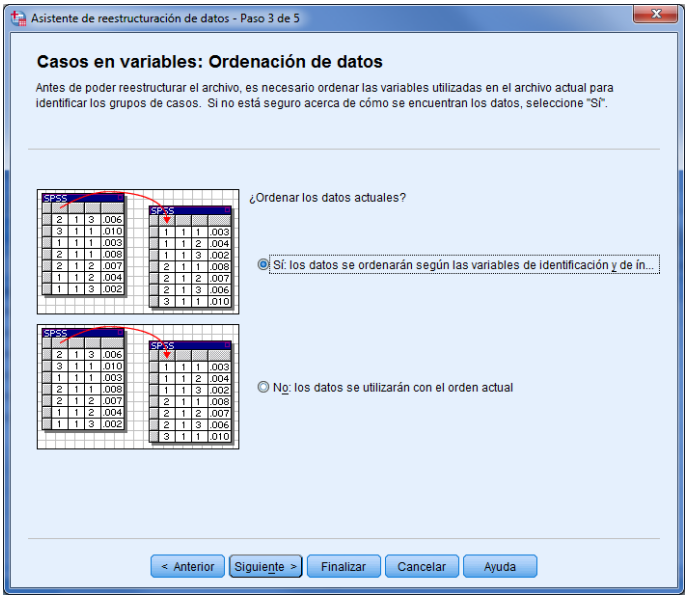
En este caso podemos estar interesados en pasar la información de las filas a las columnas, para tener 3 casos y 4 variables (el contrato y salario en los dos momentos). Para ello ejecutamos el procedimiento de reestructuración por el menú Datos / Reestructurar (comando **CASESTOVARS**) y elegimos la opción Reestructurar casos seleccionados en variables:



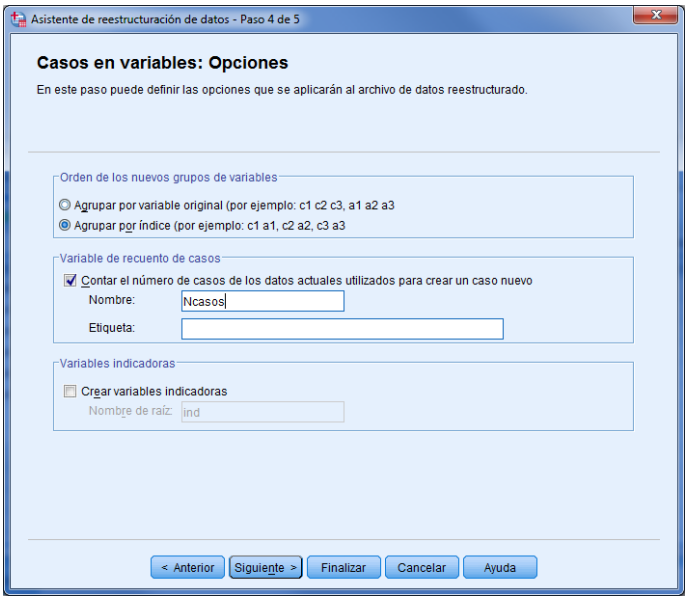
En la siguiente ventana elegimos la variable de identificación del grupo de casos, en nuestro caso ID:



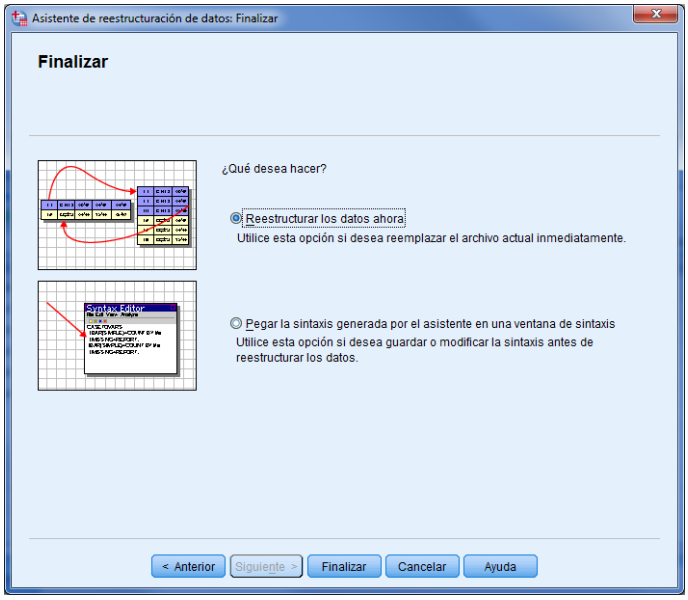
En los datos originales, una variable aparece en una única columna. En el nuevo archivo de datos, dicha variable aparecerá en varias columnas. Las variables de índice son variables existentes para crear las nuevas columnas. Los datos reestructurados contendrán una nueva variable por cada valor exclusivo contenido en dichas columnas. En este caso no las utilizamos. En el paso 3 del asistente elegiremos la opción por defecto de ordenar los datos según la variable de identificación (de hecho coincide con la actual):



En cuarto lugar decidimos cómo ordenar las variables en la nueva matriz, optamos por agrupar por índice, y calculamos una variable con el número de casos (Ncasos):



Finalmente se ejecuta el procedimiento directamente o se convierte en sintaxis:

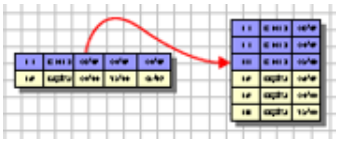


El resultado es la matriz siguiente:

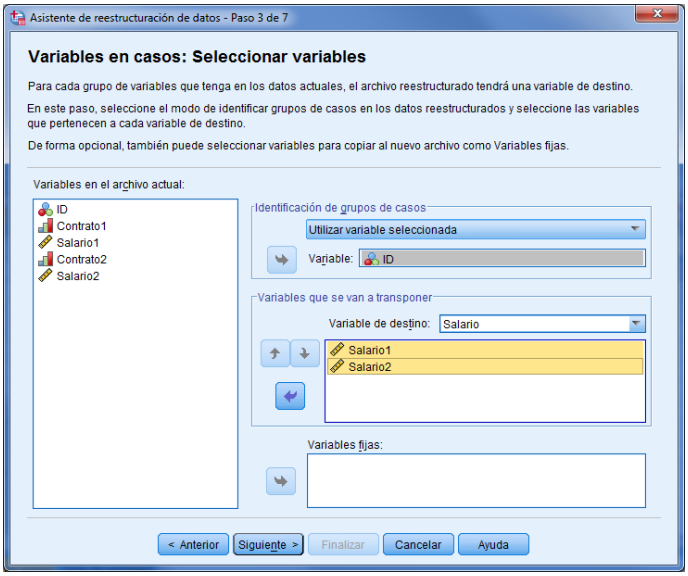
	ID	Ncasos	Momento.1	Contrato.1	Salario.1	Momento.2	Contrato.2	Salario.2
1	1	2	Inicial	Temporal	1000	Actual	Fijo	1200
2	2	2	Inicial	Fijo	1500	Actual	Temporal	1000
3	3	2	Inicial	Fijo	2000	Actual	Fijo	3000

Si nos encontramos en la situación inversa, con información en las columnas que queremos pasar a las filas, el caso de la matriz de datos casestovars.sav:

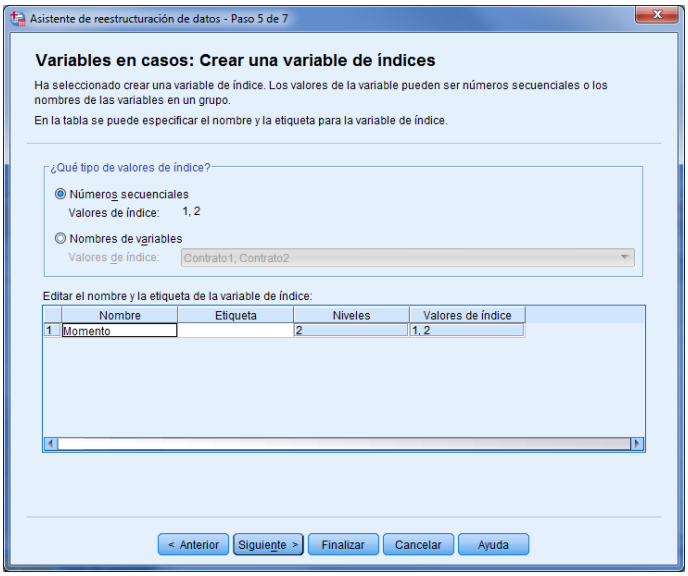
	ID	Contrato1	Salario1	Contrato2	Salario2
1	1	Temporal	1000	Fijo	1200
2	2	Fijo	1500	Temporal	1000
3	3	Fijo	2000	Fijo	3000



El proceso a seguir será similar. En este caso elegimos la opción Reestructurar variables seleccionadas en casos (comando **VARSTOCASES**), en el paso 2 elegimos reestructurar según un grupo de variables puesto que tenemos 2 variables de contrato y 2 de salario. En tercer lugar realizamos los siguientes ajustes: en la identificación de los grupos elegimos la opción Utilizar variable seleccionada y pasamos la variable ID, en la selección de las variables a transponer primero cambiamos el nombre que aparece para el primer grupo, trans1, por Contrato, y pasamos las variables Contrato1 y Contrato2; lo mismo operamos con trans2 que nombraremos como Salario y pasaremos Salario1 y Salario2:



En el cuarto paso dejamos la opción de creación de una sola variable índice. En el quinto dejamos la opción por defecto de crear números secuenciales y cambiamos el nombre de la varaible Indice1 por Momento:



En el sexto paso dejamos las opciones por defecto y clicamos sobre finalizar en el último. El resultado es una matriz de datos con esta disposición:

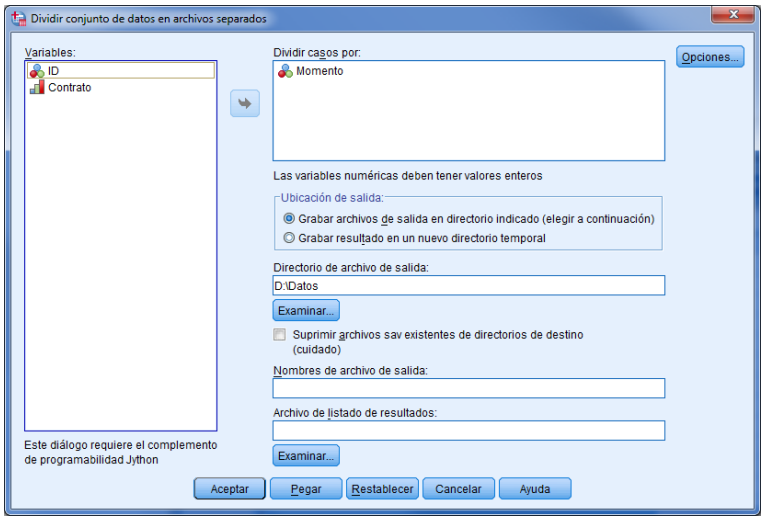
	ID	Momento	Contrato	Salario
1	1	1	Temporal	1000
2	1	2	Fijo	1200
3	2	1	Fijo	1500
4	2	2	Temporal	1000
5	3	1	Fijo	2000
6	3	2	Fijo	3000

2.1.1.2. *Tratamiento de datos entre ficheros que se relacionan*

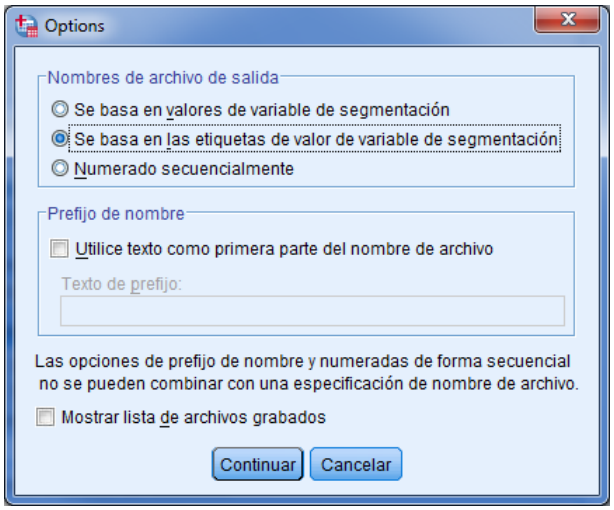
Veremos a continuación otras tareas de manipulación de matrices de datos que implican relacionar dos o más archivos: la división y la fusión.

Dividir en archivos

Es un procedimiento que actúa de forma similar a la segmentación pero su función es grabar las divisiones en nuevos archivos, de especial interés cuando necesitamos operar procedimientos distintos según el grupo de segmentación. El comando **SPSSINC PROCESS FILES** (menú Datos / Dividir en archivos) realiza esta tarea. Como ejercicio tomaremos la matriz casestovars.sav y la dividiremos entre la información del momento inicial y del momento actual. Especificamos pues que la variable de segmentación es Momento e indicamos la carpeta donde se guardarán los datos:



Completamos el procedimiento clicando sobre Opciones y elegimos que nombre los archivos de salida según las etiquetas de la varaible de segmentación.

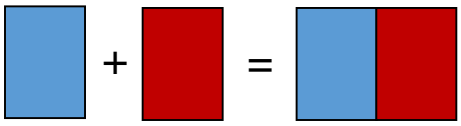


Después de darle a Continuar y Aceptar se obtienen las dos matrices: Inicial.sav y Actual.sav con tres casos cada una.

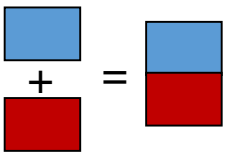
Fusionar archivos

La fusión o unión de archivos da lugar a dos alternativas:

- Añadir variables. Se fusiona el archivo de datos activo con otro que contiene los mismos casos pero variables diferentes.



- Añadir casos. Se fusiona el archivo de datos activo con otro que contiene las mismas variables pero casos diferentes.



Realizaremos un pequeño ejercicio con la matriz Y.sav que contine 6 casos y 4 variables, Edad y Sexo son características individuales sociodemográficas y Sector y Tamaño hacen referencia a características laborales de la empresa:

	ID	Edad	Sexo	Sector	Tamaño
1	1	23	Mujer	Servicios	20
2	2	35	Varón	Primario	1
3	3	48	Varón	Industria	100
4	4	55	Mujer	Industria	500
5	5	28	Varón	Construcción	50
6	6	20	Varón	Servicios	5

Para el ejercicio de unir variables consideraremos dos matrices iniciales separadas con la información sociodemográfica (YA.sav) y la información de la empresa (YB.sav). Para el ejercicio de unir casos disponemos de dos matrices separadas con los tres primeros casos (Y1-3.sav) y los tres últimos (Y4-6.sav).

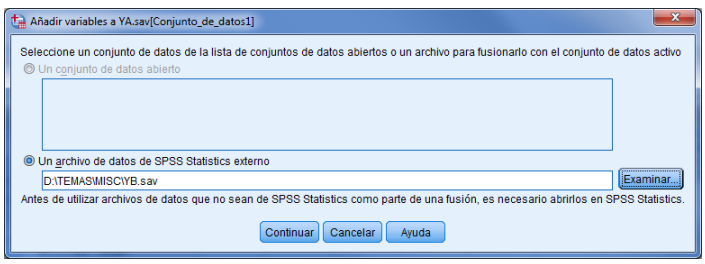
	ID	Edad	Sexo	Sector	Tamaño
1	1	23	Mujer	Servicios	20
2	2	35	Varón	Primario	1
3	3	48	Varón	Industria	100
4	4	55	Mujer	Industria	500
5	5	28	Varón	Construcción	50
6	6	20	Varón	Servicios	5

YA YB

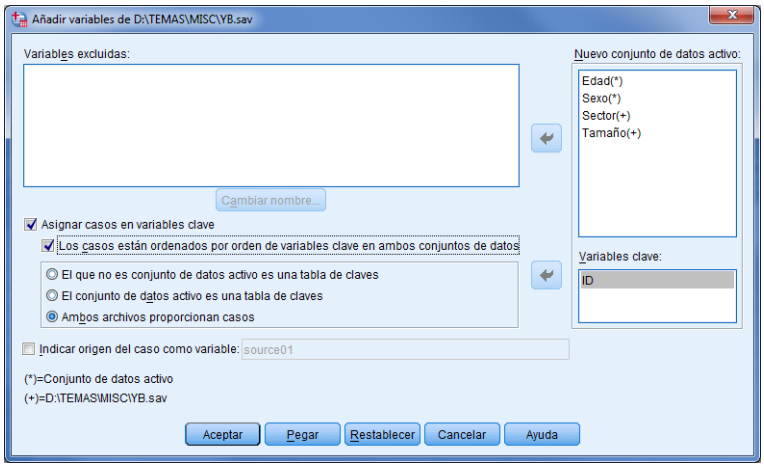
	ID	Edad	Sexo	Sector	Tamaño
1	1	23	Mujer	Servicios	20
2	2	35	Varón	Primario	1
3	3	48	Varón	Industria	100
4	4	55	Mujer	Industria	500
5	5	28	Varón	Construcción	50
6	6	20	Varón	Servicios	5

Y1-3 Y4-6

En el primer caso la fusión se realiza con el comando es **MATCH FILES** (menú Datos / Fusiona / Añadir variables). Abrimos en primer lugar la matriz YA.sav y a continuación añadimos las variables de la matriz YB.sav:



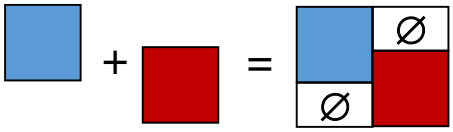
Podemos tenerla abierta y elegirla en el primer recuadro o bien ir a buscarla a la carpeta donde esté guardada. Le damos a continuar y nos aparece el cuadro de diálogo de la fusión:



Pasa fusionar es muy conveniente disponer de una variable clave que identifique a cada unidad en cada uno de los archivos a unir, de esta forma se irá emparejando la información a partir del control de la coincidencia del mismo caso. En nuestro ejemplo este papel lo juega la variable ID. Con una variable clave se requiere entonces previamente ordenar ambos ficheros por ella. El tipo de fusión que haremos implicará que Ambos archivos proporcionan casos, se trata de casos individuales en los dos archivos. Las otras dos opciones (El que no es conjunto de datos activo (o el conjunto de datos activo) es una tabla de claves) implica que existe una tabla de claves o tabla de referencia, es decir, un archivo en el que los datos de cada caso se pueden aplicar a varios casos del otro archivo de datos (una característica del hogar como atributo para todos los individuos del hogar, por ejemplo).

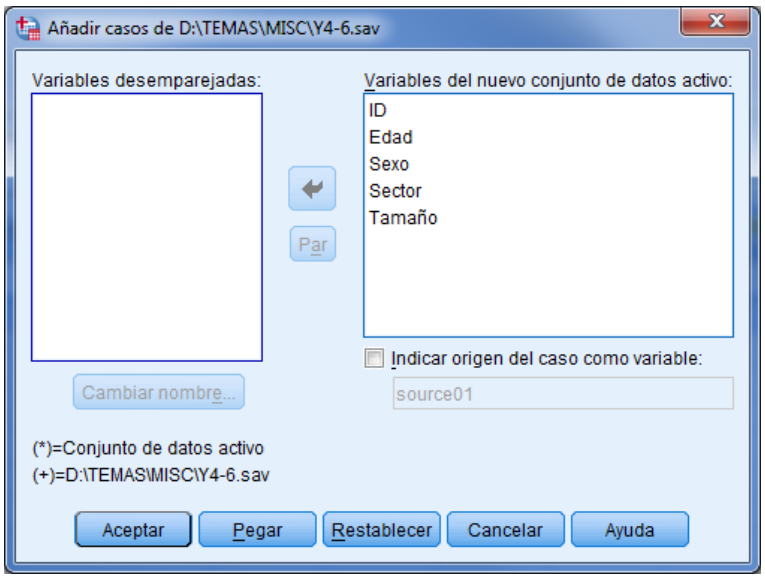
La variable ID se coloca en el recuadro Variable clave después de clicar sobre Asignar casos en variable clave. Las variables que se unen se identifican por el fichero al que pertenecen en el recuadro Nuevo conjunto de datos activo: las del fichero activo (YA.sav) con **(*)** y las del que se añade (YB.sab) con **(+)**. Las variables que son comunes del segundo archivo quedan en el recuadro Variables excluidas, donde estaba la variable ID. Una vez ejecutado tendremos como resultado la misma información de la matriz Y.

Conviene tener presente que todos los casos desemparejados, es decir, los que están en una matriz y no en la otra, sea la que sea, tendrán valores perdidos en la fusión para las variables donde no tienen información, serán vacíos (∅) en la nueva matriz:



Realizaremos ahora el segundo caso de fusión, el de añadir casos. El comando **ADD FILES** (menú Datos / Añadir casos) lo ejecutaremos a partir de la matriz Y1-3.sav a la que le añadiremos Y4-6.sav que elegiremos de la misma forma que en el caso de añadir variables. En esta ocasión nos aparecerá la lista de variables común y las variables que

quedan desemparejadas porque están en un fichero y no en el otro, éstas no se incluirán en el archivo fusionado.



.

De nuevo ejecutando el procedimiento reproducimos la matriz Y.sav.

2.1.2. *Transformación de los datos*

Después de ver distintas operaciones de tratamiento de una matriz en su conjunto nos centramos en aquellas tareas de transformación donde se implican variables concretas de la matriz, de forma individual o relacionándolas con otras. El sistema SPSS dispone de diversos comandos destinados a la transformación de las variables existentes, bien sea para su modificación o bien por la generación o creación de nuevas variables. La construcción de tipologías y de índices a partir de diversas variables será una de las necesidades frecuentes del análisis, la recodificación de los valores de las variables para agrupar valores o reducir la escala de medida es otra tarea inmediata que conlleva el análisis. Todas estas tareas se resuelven a través del menú Transformación de SPSS. Los comandos de SPSS que comentaremos son los de la Tabla III.2.3.

Tabla III.2.3 Procedimientos de transformación de variables

Menú Datos	Comandos de SPSS
Recodificar	**RECODE, AUTORECODE**
Agrupación visual	**RECODE**
Calcular	**COMPUTE**
Contar valores	**COUNT**
Calcular Si	**COMPUTE, IF**
	DO IF … END IF

En todo ejercicio de creación de variables hay que tener presente el comportamiento de los valores perdidos en dos momentos: antes y después de crear las variables. Antes, hay que tener en cuenta que si las variables contienen valores perdidos (del sistema o del usuario) en las nuevas variables estos aparecerán como valores perdidos del sistema si no se tratan específicamente. Por otra parte, cuando creamos una variable nueva

debemos prever y controlar la generación no deseada de valores perdidos como resultado de una operación en la que las transformaciones no se aplican de hecho en todos los casos que inicialmente queremos considerar. Si alguna transformación no se aplica a un caso concreto el valor de la variable creada que aparecerá será un valor perdido del sistema.

Como se trata de comandos de transformación recordemos que su ejecución no es efectiva hasta que se encuentra un comando de procedimiento que fuerce la lectura los datos del archivo (un procedmiento de análisis), función que también cumple el comando **EXECUTE**.

Hay que tener presente finalmente que toda generación de variables requiere completar su diccionario (etiquetas, formato, valores perdidos, nivel de medida, etc.) a través de la pestaña de Variables o bien a través de los comandos correspondientes de sintaxis.

2.1.2.1. *Recodificación de variables*

La recodificación de variables permite cambiar los valores actuales de las variables por otros nuevos. La recodificación puede significar estrictamente un cambio de uno o más valores por otros, o bien la combinación o la agrupación de rangos de valores en nuevas categorías. El valor a recodificar pueden ser numérico o alfanumérico (formato de cadena, *string*) y se puede pasar de una codificación alfanumérica a otra numérica.

Por otro lado la recodificación se puede realizar optando por mantener la variable original y generando una nueva con otra nombre que tendrá los valores recodificados, o bien optando por sustituir la variable que se está recodificando por la nueva variable con los nuevos criterios de codificación y con el mismo nombre de variable. El primer caso en terminología del SPSS se denomina recodificar en distintas variables y el segundo caso recodificar en las mismas variables.

El comando del SPSS que realiza la recodificación es **RECODE**. El cuadro de diálogo para efectuar la recodificación se encuentra en el menú Transformar / Recodificar donde hay que optar por la recodificación en las mismas o en distintas variables.

Nos detendremos en el segundo caso, el primero es equivalente, aunque en general conviene no utilizarlo si no se tiene la certeza de su conveniencia pues siempre implica que la variable original desaparezca. A partir de la matriz de datos CIS3041.sav realizaremos dos ejercicios de recodificación: a partir de una variable cualitativa y a partir de una cuantitativa.

El primer paso para realizar una recodificación es definir los criterios de recodificación y observar los valores de las variables extrayendo la tabla de frecuencias. Consideramos en primer lugar la variable OCUMAR11, la categoría ocupacional de la persona entrevistada según la CNO de 2011 (Clasificación Nacional de Ocupaciones)[24] . Su tabla de frecuencias es esta:

[24] La CNO (http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft40%2Fcno11%2F&file=inebase&L=0) es la adaptación española de la clasificación internacional ISCO (*International Standard Classification of Occupations*) de la OIT (http://www.ilo.org/public/spanish/bureau/stat/isco/), o CIUO, que tiene varios niveles de desagregación, hasta 5 y se codifica a 4 dígitos. Aquí se presenta con un 1 solo dígito. La variable P40 de la matriz

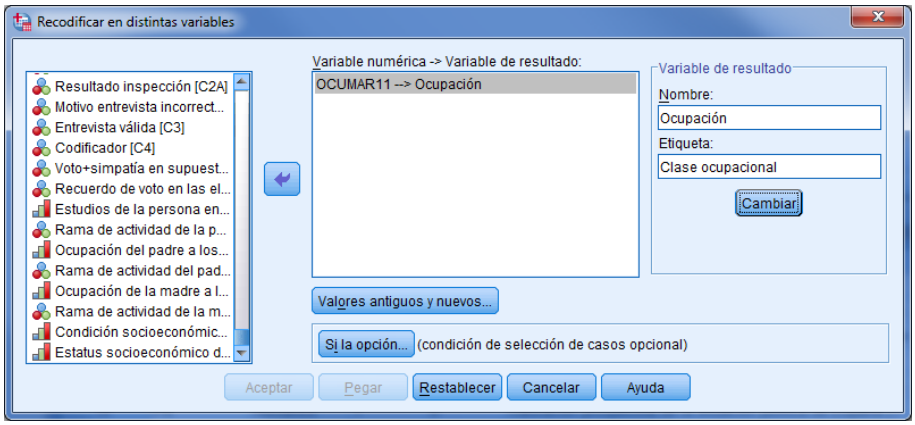
OCUMAR11 Ocupación de la persona entrevistada (CNO11)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Directores/as y gerentes	84	3,4	3,4	3,4
	2 Técnicos/as y profesionales científicos/as e intelectuales	309	12,5	12,6	16,1
	3 Técnicos/as; profesionales de apoyo	325	13,1	13,3	29,4
	4 Empleados/as contables, administrativos/as y otros empleados/as de oficina	100	4,0	4,1	33,4
	5 Trabajadores/as de los servicios de restauración, personales, protección y vendedores/as	559	22,5	22,9	56,3
	6 Trabajadores/as cualificados/as en el sector agrícola, ganadero, forestal y pesquero	132	5,3	5,4	61,7
	7 Artesanos/as y trabajadores/as cualificados/as de las industrias manufactureras y la construcción, excepto operadores/as	359	14,5	14,7	76,4
	8 Operadores/as de instalaciones y maquinaria, y montadores/as	274	11,0	11,2	87,6
	9 Ocupaciones elementales	294	11,9	12,0	99,6
	10 Ocupaciones militares	10	,4	,4	100,0
	Total	2446	98,6	100,0	
Perdidos	94 Sin ocupación/vive de las rentas	1	,0		
	98 N.S./Ocupación mal especificada o insuficiente	13	,5		
	99 N.C.	20	,8		
	Total	34	1,4		
Total		2480	100,0		

El objetivo es disponer de una variable ocupacional con un número más reducido de categorías a partir de la agrupación de las 10 que tiene la variable original. Consideraremos una agrupación en 4 categorías ocupacionales más una categoría de valores perdidos según los siguientes criterios:

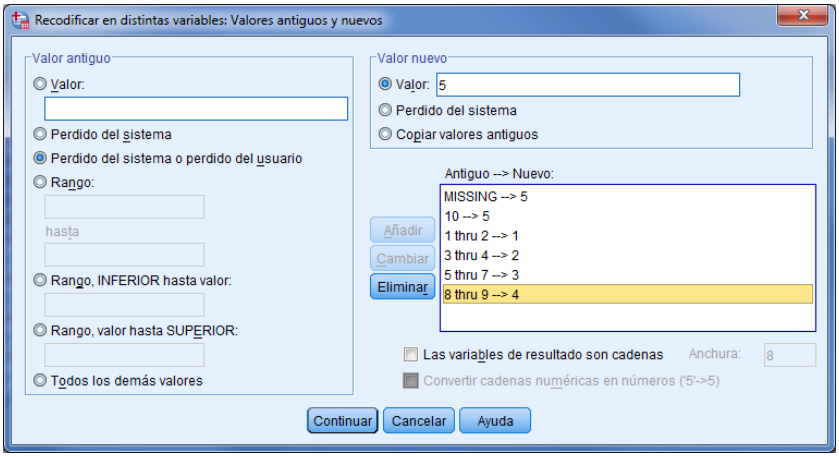
1. Clase alta o categoría ocupacional alta: códigos 1 y 2.
2. Clase media o categoría ocupacional intermedia: códigos 3 y 4.
3. Trabajadores cualificados o categoría media baja: códigos 5, 6 y 7.
4. Trabajadores no cualificados o categoría baja: códigos 8 y 9.
5. Valores perdidos: códigos 10, 94, 98 y 99.

Entramos en el menú del procedimiento y elegimos la variable OCUMAR11 para pasarla al recuadro de la derecha. Aparecerá el nombre y un **?**, para indicarnos que debemos darle nombre a la nueva variable. En el apartado de Variables de resultado escribimos el nombre de la nueva variable, por ejemplo Ocupación, y una etiqueta, Clase ocupacional en este caso. Para hacer efectiva la acción en necesario clicar sobre Cambiar:



A continuación debemos especificar la correspondencia entre los Valores antiguos y los nuevos, clicamos sobre dicho botón:

CIS3041.sav es la CNO 2011 a tres dígitos. Por tanto, de hecho OCUMAR11 es ya una variable que ha sido recodificada (agrupada) a un solo dígito.



Los criterios de recodificación que hemos comentado se trasladan de la forma siguiente: para los 4 primeros nuevos valores elegiremos la opción Rango especificando en cada caso el valor inferior y superior. El primer el rango sería 1 hasta 2, como especificación del lado izquierdo (valor antiguo), en el lado de la derecha (valor nuevo) escribiremos 1 en la casilla de Valor, y el botón Añadir a continuación. Así definimos que Directores y gerentes junto a Técnicos y profesionales, valores 1 y 2, se unan en una sola categoría, codificada con valor 1. Así seguiríamos con los tres casos siguientes como se puede ver en la imagen. El valor 10 lo consideraremos como valor perdido junto con los valores perdidos que ya tiene la variable (sin ocupación, NS, NC). Éstos corresponden a los códigos 94, 98 y 99, pero como todos están considerados valores perdidos del usuario en la variable original nos podemos referir a ellos conjuntamente como Valores perdidos del sistema o del usuario, palabra clave MISSING en SPSS). Le damos a continuar y a aceptar para ejecutar la recodificación. Para ver el resultado necesitamos pedir la tabla de frecuencias, el resultado es el siguiente:

Ocupación Clase ocupacional

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1,00	393	15,8	15,8	15,8
	2,00	425	17,1	17,1	33,0
	3,00	1050	42,3	42,3	75,3
	4,00	568	22,9	22,9	98,2
	5,00	44	1,8	1,8	100,0
	Total	2480	100,0	100,0	

Se puede comprobar cómo se corresponden las frecuencias de la nueva variable a la suma de las categorías de la variable original. En la tabla vemos los nuevos valores pero no tienen etiquetas. Como sugerimos, después de la creación de una variable es preciso completar su diccionario. Es necesario poner las etiquetas de los valores, precisar que no tiene decimales, definir el 5 como valor perdido del usuario y poner su nivel de medición como ordinal. Volvemos a pedir la tabla y el resultado final es estos arreglos:

Ocupación Clase ocupacional

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Clase alta	393	15,8	16,1	16,1
	2 Clase media	425	17,1	17,4	33,6
	3 Trababajadores cualificados	1050	42,3	43,1	76,7
	4 Trabajadores no cualificados	568	22,9	23,3	100,0
	Total	2436	98,2	100,0	
Perdidos	5 Perdidos:NS,NC,FFAA	44	1,8		
Total		2480	100,0		

► Ejercicio 5. Propuesto

El INE en su informe *Introducción a la CNO-11* comenta el concepto de ocupación (http://www.ine.es/daco/daco42/clasificaciones/Introduccion_CNO11.V02.pdf) y distingue entre empleo y competencias como dos dimensiones fundamentales que lo estructuran. Las competencias distinguen a su vez dos dimensiones: la especialización y el nivel de competencias. Éste último tiene 4 grados (asociados teóricamente a los niveles educativos formales) que se corresponden con las categorías ocupacionales a 1 dígito de la forma siguiente:

Cuadro 1: Correspondencia entre los Grandes Grupos de la CIUO-08 y el nivel de competencias

Grandes Grupos CIUO-08	Nivel de competencias
1 - Directores y gerentes	3, 4
2 - Profesionales científicos e intelectuales	4
3 - Técnicos y profesionales de nivel medio	3
4 - Personal de apoyo administrativo 5 - Trabajadores de los servicios y vendedores de comercios y mercados 6 - Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros 7 - Oficiales, operarios y artesanos de artes mecánicas y de otros oficios 8 - Operadores de instalaciones y máquinas y ensambladores	2
9- Ocupaciones elementales	1
0- Ocupaciones militares	1, 2, 4

Fuente: INE

De acuerdo con esta tabla, sin considerar el 0 de las ocupaciones militares y asignando a los directores y gerentes sólo el nivel 4, agrupar los grandes grupos ocupacionales (variable OCUMAR11 de la matriz CIS3041.sav) en los 4 niveles de competencias. Completar igualmente el diccionario de las variables y extraer la tabla de frecuencias para comprobar el resultado.

Un segundo ejemplo de recodificación tendrá en cuenta una variable cuantitativa, la edad (variable P32). Es habitual trabajar con la edad agrupada en intervalos de 5 o 10 años, o en grandes grupos de edad (jóvenes, adultos, mayores). Así la variable original cuantitativa reduce su escala y permite trabajarla con menos categorías como una variable cualitativa ordinal. Se propone crear una nueva variable de edad (Edad10) con una agrupación en intervalos según estos criterios:

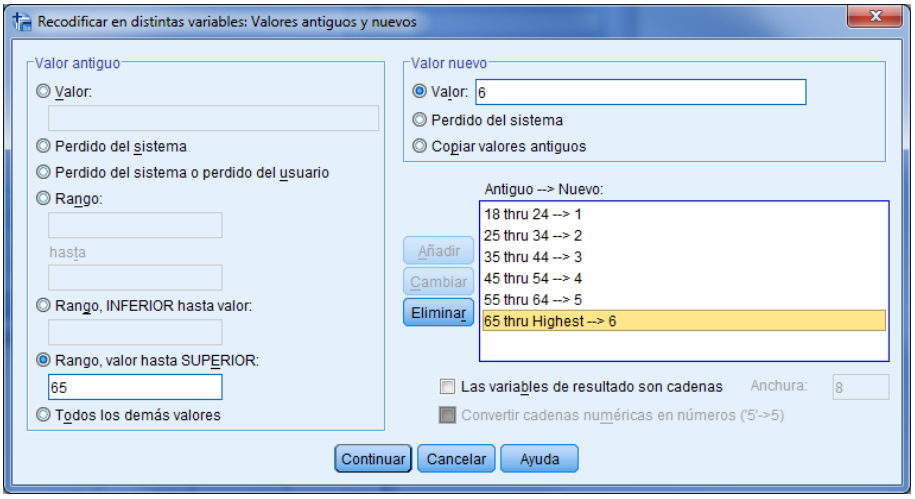
1. 18 a 24
2. 25 a 34
3. 35 a 44
4. 45 a 54
5. 55 a 64
6. 65 y más

Como la variable original no tiene valores perdidos no es necesario contemplarlos en la nueva. La tabla de distribución de frecuencias original es la siguiente:

P32 Edad de la persona entrevistada

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	18	32	1,3	1,3	1,3
	19	32	1,3	1,3	2,6
	20	18	,7	,7	3,3
	21	28	1,1	1,1	4,4
	22	38	1,5	1,5	6,0
	23	25	1,0	1,0	7,0
	24	27	1,1	1,1	8,1
	25	51	2,1	2,1	10,1
	26	42	1,7	1,7	11,8
	27	40	1,6	1,6	13,4
	28	23	,9	,9	14,4
	29	39	1,6	1,6	15,9
	30	46	1,9	1,9	17,8
	31	48	1,9	1,9	19,7
	32	41	1,7	1,7	21,4
	33	47	1,9	1,9	23,3
	34	53	2,1	2,1	25,4
	35	51	2,1	2,1	27,5
	36	37	1,5	1,5	29,0
	37	48	1,9	1,9	30,9
	38	47	1,9	1,9	32,8
	39	46	1,9	1,9	34,6
	40	48	1,9	1,9	36,6
	41	43	1,7	1,7	38,3
	42	57	2,3	2,3	40,6
	43	61	2,5	2,5	43,1
	44	71	2,9	2,9	45,9
	45	51	2,1	2,1	48,0
	46	51	2,1	2,1	50,0
	47	45	1,8	1,8	51,9
	48	42	1,7	1,7	53,5
	49	45	1,8	1,8	55,4
	50	57	2,3	2,3	57,7
	51	33	1,3	1,3	59,0
	52	34	1,4	1,4	60,4
	53	49	2,0	2,0	62,3
	54	56	2,3	2,3	64,6
	55	34	1,4	1,4	66,0
	56	36	1,5	1,5	67,4
	57	42	1,7	1,7	69,1
	58	29	1,2	1,2	70,3
	59	27	1,1	1,1	71,4
	60	48	1,9	1,9	73,3
	61	27	1,1	1,1	74,4
	62	33	1,3	1,3	75,7
	63	39	1,6	1,6	77,3
	64	40	1,6	1,6	78,9
	65	48	1,9	1,9	80,8
	66	37	1,5	1,5	82,3
	67	39	1,6	1,6	83,9
	68	24	1,0	1,0	84,9
	69	31	1,3	1,3	86,1
	70	36	1,5	1,5	87,6
	71	27	1,1	1,1	88,7
	72	28	1,1	1,1	89,8
	73	18	,7	,7	90,5
	74	21	,8	,8	91,4
	75	19	,8	,8	92,1
	76	20	,8	,8	92,9
	77	18	,7	,7	93,7
	78	25	1,0	1,0	94,7
	79	16	,6	,6	95,3
	80	17	,7	,7	96,0
	81	17	,7	,7	96,7
	82	14	,6	,6	97,3
	83	15	,6	,6	97,9
	84	13	,5	,5	98,4
	85	11	,4	,4	98,8
	86	8	,3	,3	99,2
	87	4	,2	,2	99,3
	88	5	,2	,2	99,5
	89	4	,2	,2	99,7
	90	2	,1	,1	99,8
	91	2	,1	,1	99,8
	92	1	,0	,0	99,9
	94	3	,1	,1	100,0
	Total	2480	100,0	100,0	

Siguiendo el protocolo que vimos anteriormente especificaremos en particular los criterios de recodificación:



La tabla de frecuencias resultante después de completar el diccionario de los datos es la siguiente:

Edad10 Edad de la persona entrevistada en grupos de 10

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 18-24	200	8,1	8,1	8,1
	2 25-34	430	17,3	17,3	25,4
	3 35-44	509	20,5	20,5	45,9
	4 45-54	463	18,7	18,7	64,6
	5 55-64	355	14,3	14,3	78,9
	6 65 y más	523	21,1	21,1	100,0
	Total	2480	100,0	100,0	

► Ejercicio 6. Propuesto

Recodificar la variable P15 de autoposicionamiento ideológico en tres categorías que agrupen los valores 1 a 3, 4 a 6 i 7 a 10.

Por otro lado, si con los datos de la encuesta del CIS nos preguntamos ¿cuáles son los ingresos medios de los hogares de los entrevistados? Para responder a esta pregunta deberíamos tener la variable de ingresos como cuantitativa y en la encuesta se pregunta por intervalos de forma cualitativa. Una alternativa es calcular la media a partir de la marca de clase de cada intervalo para lo que deberemos recodificar la variable. La distribución de la variable de ingresos (P45) es la siguiente:

P45 Ingresos del hogar

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 No tienen ingresos de ningún tipo	10	,4	,6	,6
	2 Menos o igual a 300 €	28	1,1	1,6	2,2
	3 De 301 a 600 €	185	7,5	10,8	13,1
	4 De 601 a 900 €	297	12,0	17,4	30,5
	5 De 901 a 1.200 €	347	14,0	20,3	50,8
	6 De 1.201 a 1.800 €	386	15,6	22,6	73,4
	7 De 1.801 a 2.400 €	215	8,7	12,6	86,0
	8 De 2.401 a 3.000 €	120	4,8	7,0	93,1
	9 De 3.001 a 4.500 €	76	3,1	4,5	97,5
	10 De 4.501 a 6.000 €	31	1,3	1,8	99,4
	11 Más de 6.000 €	11	,4	,6	100,0
	Total	1706	68,8	100,0	
Perdidos	99 N.C.	774	31,2		
Total		2480	100,0		

Si recodificamos a través de la sintaxis de SPSS llamando a la nueva variable P45m podemos utilizar las instrucciones siguientes que contemplan, además de la recodificación, completar el diccionario de la variable y el cálculo de las frecuencias junto al estadístico de la media:

```
FREQUENCIES P45.
RECODE P45 (1=0)(2=150)(3=450)(4=750)(5=1050)(6=1500)(7=2100)(8=2700)
          (9=3750)(10=5250)(11=7500)(MISSING=9999)  INTO P45m.
VARIABLE LABELS  P15m 'Ingresos del hogar (marca de clase)'.
VALUE LABELS P45m 9999 'NC'.
MISSING VALUES P45m(9999).
FORMATS  P45m (F2.0).
VARIABLE LEVEL  P45m (SCALE).
FREQUENCIES P45m /STATISTICS MEAN.
```

Este es el resultado:

P45m Ingresos del hogar (marca de clase)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0	10	,4	,6	,6
	150	28	1,1	1,6	2,2
	450	185	7,5	10,8	13,1
	750	297	12,0	17,4	30,5
	1050	347	14,0	20,3	50,8
	1500	386	15,6	22,6	73,4
	2100	215	8,7	12,6	86,0
	2700	120	4,8	7,0	93,1
	3750	76	3,1	4,5	97,5
	5250	31	1,3	1,8	99,4
	7500	11	,4	,6	100,0
	Total	1706	68,8	100,0	
Perdidos	9999 NC	774	31,2		
Total		2480	100,0		

Estadísticos

P45m Ingresos del hogar (marca

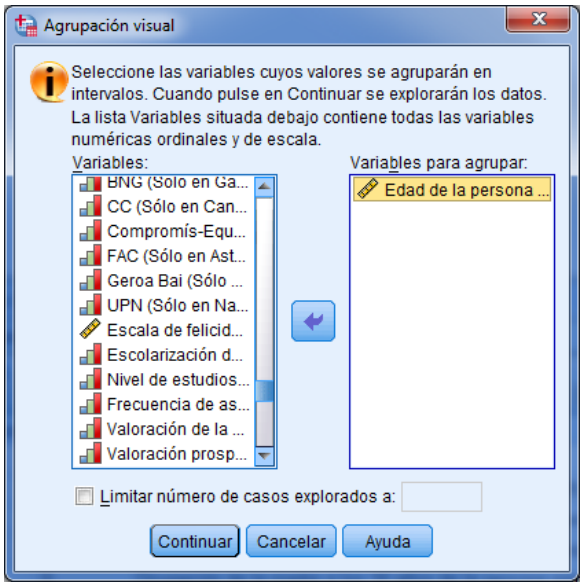
N	Válido	1706
	Perdidos	774
Media		1500,18

La media de los ingresos de los hogares de la muestra es de 1500 €.

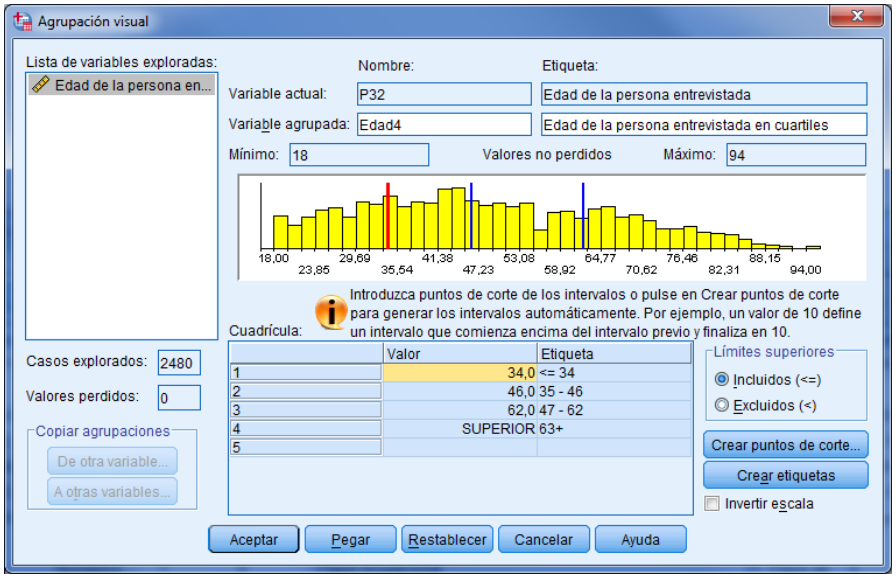
► Ejercicio 7. Propuesto

Recodificar la variable P46 relativa a los ingresos personales con la marca de clase de los intervalos y calcular la media de los ingresos.

En SPSS se dispone de un interesante procedimiento asistido y automatizado de recodificación de variables cuantitativas denominado Agrupación visual en el menú Transformar. Cuando se accede debemos elegir en primer lugar la variable, podemos elegir la P32 de la edad:

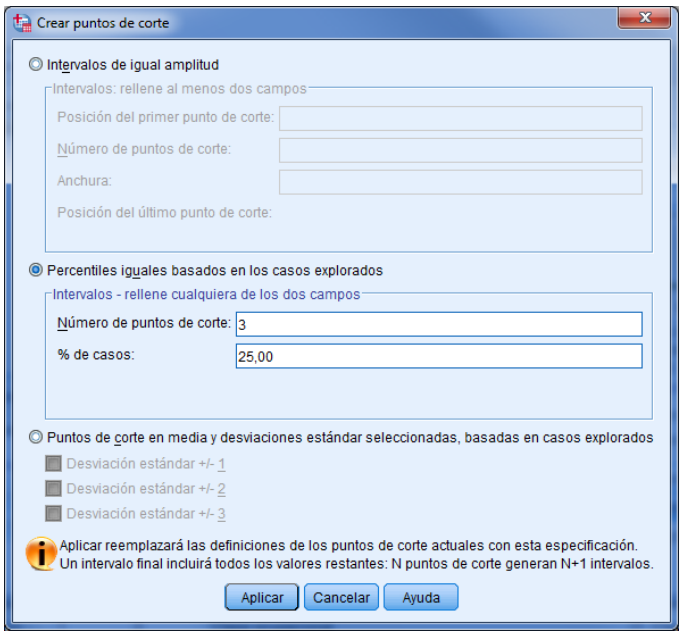


Después de darle a continuar accedemos a este cuadro dialogo donde ya hemos especificado las distintas opciones que ahora comentamos:



Inicialmente nos aparece el histograma sin particiones o agrupaciones de los valores y con una propuesta de etiqueta para la variable que se crea. Nos informa también de los valores mínimo (18) y máximo (94). Debemos dar un nombre a la nueva variable, por ejemplo Edad4. En la parte inferior se detallan los criterios de recodificación y las etiquetas. Podemos escoger los puntos de corte manualmente, poniendo los valores correspondientes en la tabla, o podemos hacerlo mediante un proceso automatizado con diversas alternativas en la pestaña Crear puntos de corte. Si optamos por esta última alternativa, en la nueva ventana de diálogo podemos escoger tres opciones:

- Intervalos de igual amplitud según el número o la anchura.
- Percentiles iguales según el número de cortes o el porcentaje de casos.
- Puntos de corte a partir de la media y las desviaciones típicas.



Cualquier alternativa podría ser válida, en este caso elegiremos crear una división de los valores de la variable en cuartiles, en 4 grupos con el 25% de los casos, lo que implica especificar 3 puntos de corte (recordemos que los cuartiles son 3, los 3 valores

que marcan los cortes). Clicamos en aceptar y al volver al cuadro de diálogo anterior clicaremos en Crear etiquetas y nos las creará de forma automática en correspondencia con los valores de la división en cuartiles. Tras ejecutar el procedimiento de recodificación y pedir la tabla de frecuencias obtenemos este resultado:

Edad4 Edad de la persona entrevistada en cuartiles

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 <= 34	630	25,4	25,4	25,4
	2 35 - 46	611	24,6	24,6	50,0
	3 47 - 62	637	25,7	25,7	75,7
	4 63+	602	24,3	24,3	100,0
	Total	2480	100,0	100,0	

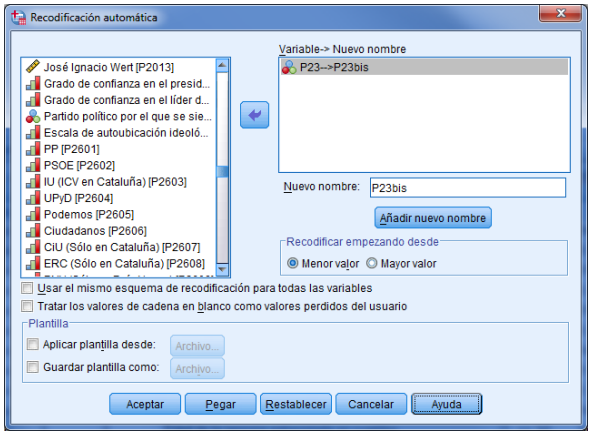
► Ejercicio 8. Propuesto

Con la matriz de datos IDH2014.sav realizar una recodificación de la variable GDPpercapita (*Gross Domestic Product per capita*) siguiendo diversos criterios: agrupar en intervalos de igual amplitud, en percentiles o a partir de unidades de desviación.

Además de la recodificación que se opera con el comando **RECODE** existe otro de recodificación automática denominado **AUTORECODE** que convierte los valores numéricos y de cadena en valores enteros consecutivos. Esta recodificación es interesante ya que algunos procedimientos de análisis no pueden utilizar variables en formato de cadena y otros requieren obligatoriamente el tratamiento de valores enteros consecutivos. También es de interés para exportar datos a otro software que trabaja las variables cualitativas con valores enteros consecutivos.

La nueva variable generada por la recodificación automática conserva las etiquetas de los valores de la variable original; en el caso de que los valores no tengan una etiqueta de valor definida se utilizará el valor original como etiqueta del valor recodificado. Cuando se trata de valores de cadena se recodifican por orden alfabético, y primero las mayúsculas antes de las minúsculas. A los valores perdidos se le asignan los últimos números consecutivos. Cuando se ejecuta el procedimiento una tabla muestra la correspondencia entre los valores antiguos, los nuevos y las etiquetas.

Por ejemplo, si quisiéramos crear códigos consecutivos para la variable P23 de intención de voto, a través del procedimiento Transformar / Recodificación automática sencillamente elegiríamos la variable original P23, le daríamos nombre a la nueva, P23bis por ejemplo, y ejecutaríamos:



En las tablas de frecuencia siguientes se pueden comparar los efectos del cambio:

P23 Intención de voto en supuestas elecciones generales

		Frecuencia	Porcentaje válido	Porcentaje acumulado
Válido	1 PP	290	11,7	11,7
	2 PSOE	354	14,3	26,0
	3 IU (ICV en Cataluña)	91	3,7	29,6
	4 UPyD	53	2,1	31,8
	5 CiU	49	2,0	33,8
	6 Amaiur	9	,4	34,1
	7 PNV	11	,4	34,6
	8 ERC	48	1,9	36,5
	9 BNG	3	,1	36,6
	10 CC	3	,1	36,7
	11 Compromís-Equo	8	,3	37,1
	12 FAC	2	,1	37,1
	13 Geroa Bai	1	,0	37,2
	14 UPN	2	,1	37,3
	15 Podemos	437	17,6	54,9
	16 Ciudadanos	37	1,5	56,4
	17 Otros partidos	44	1,8	58,1
	77 Voto nulo	2	,1	58,2
	96 En blanco	105	4,2	62,5
	97 No votaría	389	15,7	78,1
	98 No sabe todavía	483	19,5	97,6
	99 N.C.	59	2,4	100,0
	Total	2480	100,0	

P23bis Intención de voto en supuestas elecciones generales

		Frecuencia	Porcentaje válido	Porcentaje acumulado
Válido	1 PP	290	11,7	11,7
	2 PSOE	354	14,3	26,0
	3 IU (ICV en Cataluña)	91	3,7	29,6
	4 UPyD	53	2,1	31,8
	5 CiU	49	2,0	33,8
	6 Amaiur	9	,4	34,1
	7 PNV	11	,4	34,6
	8 ERC	48	1,9	36,5
	9 BNG	3	,1	36,6
	10 CC	3	,1	36,7
	11 Compromís-Equo	8	,3	37,1
	12 FAC	2	,1	37,1
	13 Geroa Bai	1	,0	37,2
	14 UPN	2	,1	37,3
	15 Podemos	437	17,6	54,9
	16 Ciudadanos	37	1,5	56,4
	17 Otros partidos	44	1,8	58,1
	18 Voto nulo	2	,1	58,2
	19 En blanco	105	4,2	62,5
	20 No votaría	389	15,7	78,1
	21 No sabe todavía	483	19,5	97,6
	22 N.C.	59	2,4	100,0
	Total	2480	100,0	

2.1.2.2. *Expresiones de transformación*

Veremos a continuación los procedimientos de transformación que implican la realización de un cálculo o una transformación condicional para generar nuevas variables. La utilización de sus comandos implica trabajar con las llamadas expresiones de transformación que se especifican en la sintaxis de las instrucciones de los comandos de transformación utilizando diferentes tipos de operadores y funciones. Existen tres tipos de expresiones: numéricas, alfanuméricas (cadena) y lógicas.

Las expresiones numéricas se emplean para crear nuevas variables numéricas y en donde se utilizan:

- Operadores aritméticos: +, –, *, /, **. Se utilizan para variables numéricas, no pueden aparecer dos seguidos y no pueden introducirse antes o después de un operador lógico o relacional. Se ejecutan después de las funciones, y al mismo nivel se ejecutan de izquierda a derecha.
- Constantes numéricas (valores numéricos).
- Funciones numéricas: son funciones que devuelven siempre un número (o un valor perdido del sistema). Se especifican a través de uno o más argumentos entre paréntesis. Pueden incluir operadores aritméticos, constantes y variables. Por ejemplo, **MEAN(V1,V2)**, calcula para cada individuo la media de dos variables. Tipos de funciones numéricas:
    - Funciones aritméticas: ABS, RND, TRUNC, SQRT, EXP, LG10, LN.
    - Funciones estadísticas: MEAN, MEDIAN, SD, VARIANCE, MIN, MAX, CFVAR.
    - Funciones de variable aleatoria y funciones de distribución: las funciones CDF, PDF, RV, SIG, IDF, NCDF, NPDF son prefijos de las distribuciones (sufijos) NORMAL, LOGISTIC, CHISQ, POISSON, F, T, BINOM, etc.
    - Funciones de fecha y tiempo: DATE, TIME, CTIME, YRMODA, XDATE, DATEDIFF, DATESUM.

Las expresiones alfanuméricas (*string*) se emplean con variables cadena, contantes (texto) ente comillas y funciones cadena: CHAR.INDEX, CHAR.LENGTH, CONCAT, LTRIM, VALUELABEL, etc.

Las expresiones lógicas son expresiones de transformación que se evalúan como verdaderas (valor 1) o falsas (valor 0) o como valores perdidos del sistema, a partir de condiciones establecidas sobre los datos utilizando variables, constantes, funciones, operadores relacionales y operadores lógicos. En general es aconsejable sino necesario utilizar los paréntesis para construir las expresiones.

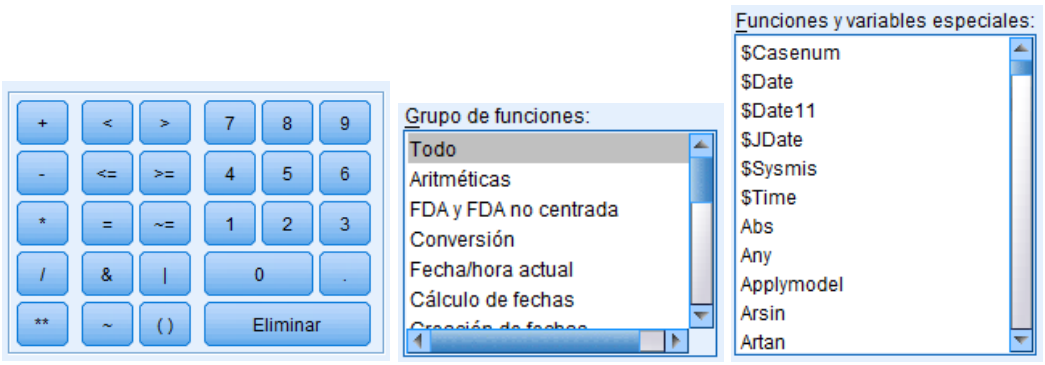
- Operadores relacionales: EQ, LT, GT, NE, LE, GE
  o bien = < > <> <= >=
- Operadores lógicos: AND, OR, NOT
  o bien & | ~
- Funciones lógicas: RANGE, ANY.

En las expresiones se evalúan primero las funciones y los operadores aritméticos, luego los operadores relacionales y los lógicos (en el orden NOT, AND, OR).

Otras funciones disponibles en SPSS son:

- Funciones de valores-missing: VALUE, MISSING, SYSMIS, NMISS, NVALID.
- Funciones de caso anterior: LAG.
- Funciones de conversión Cadena/Numérico: STRING, NUMERIC.

Cuando trabajamos por menús para construir expresiones de transformación disponemos de un asistente para recordarnos los distintos operadores y funciones como veremos a continuación.



2.1.2.3. *Cálculo de variables*

La creación de nuevas variables realizando cálculos es una necesidad constante de todo proceso de análisis de datos cuantitativos. Ya sea para modificar o combinar las variables originales existentes podemos operar infinidad de transformaciones ya sea de naturaleza estadística para acondicionar variables en un análisis, para crear indicadores y nuevas variables variables cuantitativas, para emplear variables instrumentales, etc.

El comando COMPUTE (menú Transformar / Calcular variable) está destinado a esta labor. El formato genérico de este procedimiento es:

COMPUTE variable de destino = expresión

Dentro de la expresión se pueden utilizar variables numéricas, constantes, operadores aritméticos, funciones numéricas, funciones de valores *missing*, funciones de números aleatorios y la función de fecha. Para variables alfanuméricas sólo es permitido crear una variable con un valor alfanumérico constante o copiar una variable en otra idéntica. En función de la expresión la instrucción puede ocupar tan solo una línea o diversas líneas.

Realizaremos algunos ejercicios de cálculo de variables. En primer lugar podemos plantearnos crear un índice de activismo sociopolítico a partir de las respuestas a la pregunta P14:

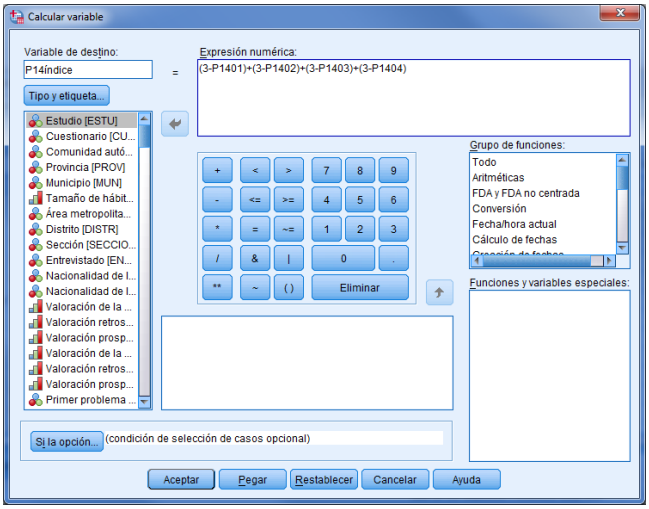
P.14 Existen diversas formas de participación en acciones sociales y políticas que la gente puede llevar a cabo. Por favor, indíqueme para cada una de ellas, si Ud.: ***(MOSTRAR TARJETA D)***.

1. Ha participado durante los últimos doce meses
2. Participó en un pasado más lejano
3. Nunca ha participado

	1	2	3	NC
- Asistir a una manifestación	1	2	3	9 (86)
- Participar en una huelga	1	2	3	9 (87)
- Participar en un foro o en un blog de debate político en Internet	1	2	3	9 (88)
- Firmar una petición /recogida de firmas bien en persona o por Internet	1	2	3	9 (89)

Con los siguientes criterios: puntuar cada forma con 2 si se ha participado últimamente, con 1 si participó en el pasado y con 0 si nunca ha participado. El índice lo construimos con esas puntuaciones en las 4 preguntas sumándolas para cada individuo. El que participe actualmente en todo tendrá un nivel de participación de 8 y el que nunca haya participado en nada de 0. A la nueva variable la llamaremos P14índice.

Teniendo en cuenta los valores actuales de la variable, la puntuación propuesta implica que antes de sumar tendremos que restar cada valor de 3 (3-1 dará 2, 3-2 dará 1 y 3-3 dará 0). Para obtener la nueva variable iremos al menú Transformar / Calcular variable. En el cuadro de diálogo pondremos el nombre de la nueva variable (P14índice) y como expresión numérica la siguiente: (3-P1401)+(3-P1402)+(3-P1403)+(3-P1404). Podemos escribir esta expresión directamente sobre el recuadro expresión numérica o podemos ayudarnos de la información disponible: las variables a la izquierda y los números, símbolos y operadores clicarlos desde los botones de la "calculadora":



Si le damos a aceptar se crea la variable. Nuestra matriz contendrá una variable más, la última. Hay que tener en cuenta que en la nueva variable algunos individuos son valores perdidos en alguna de las cuatro variables iniciales por lo que no se podrá realizar el cálculo para ellos y serán valores perdidos del sistema en la nueva[25]. Necesita completarse su diccionario (tipo, etiqueta de la variable, nivel de medición) que parcialmente podemos realizar a través del botón Tipo y etiqueta del cuadro de diálogo de Calcular. Una vez realizada la tarea la tabla de frecuencias de la nueva variable es la siguiente:

P14índice Índice de participación sociopolítica

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0	805	32,5	33,0	33,0
	1	324	13,1	13,3	46,2
	2	417	16,8	17,1	63,3
	3	299	12,1	12,2	75,5
	4	264	10,6	10,8	86,3
	5	127	5,1	5,2	91,5
	6	127	5,1	5,2	96,7
	7	36	1,5	1,5	98,2
	8	44	1,8	1,8	100,0
	Total	2443	98,5	100,0	
Perdidos	Sistema	37	1,5		
Total		2480	100,0		

Si calculamos la media se obtiene un valor de 2,09, mucho más cerca de 0 que de 8, indicando un nivel de activismo sociopolítico de la sociedad española en su conjunto relativamente bajo.

► Ejercicio 9. Propuesto

A partir de la pregunta P11 sobre la frecuencia con que se consultan los periódicos, la radio y la televisión para seguir la actualidad política, dando entre 4 y 0 puntos a las frecuencias que van de 1 (Todos los días) a 5 (Nunca) y sumando las puntuaciones para cada individuo.

Otra operación importante es la tipificación o estandarización de una variable, transformación que consiste en restar la media a cada puntuación o valor de una variable cuantitativa y dividir por la desviación típica.

$$z_i = \frac{x_i - \bar{x}}{s}$$

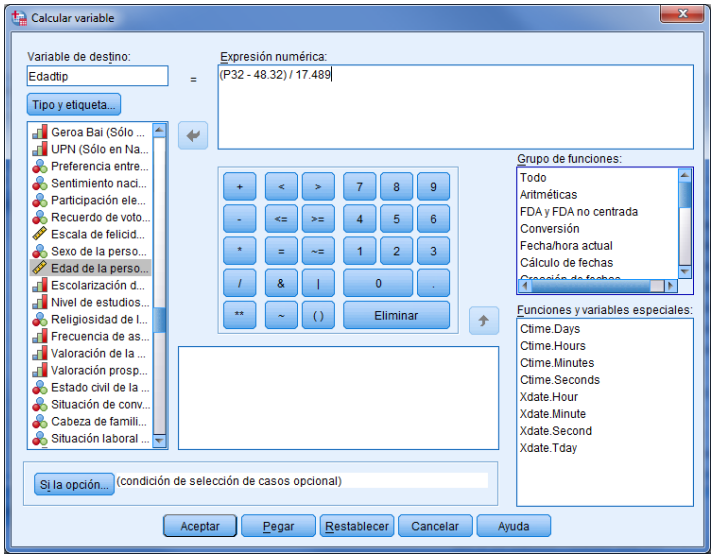
Realizamos esta operación con la variable edad (P32). Necesitamos conocer previamente los valores de la media y la desviación ejecutamos el procedimiento Analizar / Estadísticos descriptivos / Descriptivos y se obtiene:

Estadísticos descriptivos

	N	Media	Desviación estándar
P32 Edad de la persona entrevistada	2480	48,32	17,489
N válido (por lista)	2480		

[25] Si lo deseamos podemos recodificarlos a un valor determinado, etiquetarlo y declararlo valor perdido del usuario, no cambia nada, simplemente es una forma de tenerlos controlados e identificados.

Una vez conocidos los valores de la media y la desviación típica creamos la nueva variable mediante el menú Transformar / Calcular variable. Elegimos un nombre para la nueva variable, por ejemplo, Edadtip, y aplicamos la fórmula que nos da las puntuaciones tipificadas:



Si pedimos los descriptivos de la nueva variable podemos comprobar como, salvo decimales, la media es 0 y las desviación típica es 1[26].

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desviación estándar
Edadtip	2480	-1,73	2,61	-,0002	1,00000
N válido (por lista)	2480				

Procederemos ahora a la construcción de los indicadores sobre la situación política que elabora el CIS en el Barómetro[27]. Las preguntas de los barómetros de todos los meses relativas a la situación política que se utilizan en la construcción del indicador son la P4 y la P6:

P.4 Y refiriéndonos ahora a la situación política general de España, ¿cómo la calificaría Ud.: muy buena, buena, regular, mala o muy mala?

- Muy buena 1
- Buena 2
- Regular 3
- Mala..... 4 (35)
- Muy mala 5
- N.S. 8
- N.C. 9

P.6 Y, ¿cree Ud. que dentro de un año la situación política del país será mejor, igual o peor que ahora?

- Mejor 1
- Igual 2
- Peor 3 (37)
- N.S. 8
- N.C. 9

[26] Este mismo cálculo se puede obtener con SPSS a través de Analizar / Estadísticos descriptivos / Descriptivos marcando la opción Guardar valores estandarizados como variables. Si lo hacemos de la edad creará la variable zP32.

[27] Se puede consultar la metodología para la construcción de indicadores del Barómetro del CIS en la página: http://www.cis.es/cis/opencms/ES/11_barometros/metodologia.html.

El Indicador de la Situación Política Actual (SPA), a partir de la pregunta P4 se define como:

$$SPA = \frac{100.p_1 + 75.p_2 + 50.p_3 + 25.p_4 + 0.p_5}{p_1 + p_2 + p_3 + p_4 + p_5}$$

donde p_1, p_2, p_3, p_4 y p_5 son, respectivamente, los porcentajes de respuesta de las opciones muy buena, buena, regular, mala y muy mala.

El Indicador de Expectativas Políticas (IEP) a partir de la pregunta P6 será:

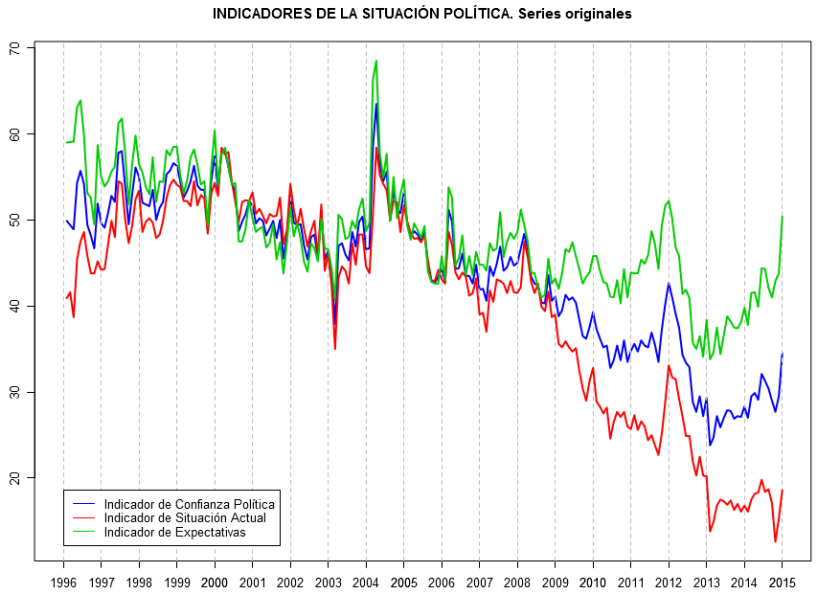
$$IEP = \frac{100.p_1 + 50.p_2 + 0.p_3}{p_1 + p_2 + p_3}$$

donde p_1, p_2 y p_3 son, respectivamente, los porcentajes de respuesta de las opciones mejor, igual y peor.

Por último, el Indicador de Confianza Política (*ICP*) es la media aritmética de los dos anteriores:

$$ICP = \frac{SPA + IEP}{2}$$

En este caso se trata de indicadores sintéticos que se expresan en un solo valor para el conjunto de la muestra, para después ser comparado a lo largo del tiempo con Barómetros anteriores[28].



Fuente: CIS

Las frecuencias de ambas variables para octubre de 2014 son:

[28] Ver http://www.cis.es/cis/export/sites/default/-Archivos/Indicadores/documentos_html/IndiPol.html.

P4 Valoración de la situación política general de España

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Muy buena	2	,1	,1	,1
	2 Buena	49	2,0	2,0	2,1
	3 Regular	357	14,4	14,9	17,0
	4 Mala	769	31,0	32,0	49,0
	5 Muy mala	1227	49,5	51,0	100,0
	Total	2404	96,9	100,0	
Perdidos	8 N.S.	59	2,4		
	9 N.C.	17	,7		
	Total	76	3,1		
Total		2480	100,0		

P6 Valoración prospectiva de la situación política de España (1 año)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Mejor	287	11,6	13,3	13,3
	2 Igual	1194	48,1	55,4	68,7
	3 Peor	676	27,3	31,3	100,0
	Total	2157	87,0	100,0	
Perdidos	8 N.S.	299	12,1		
	9 N.C.	24	1,0		
	Total	323	13,0		
Total		2480	100,0		

Para obtener los 3 indicadores utilizaremos el SPSS como "calculadora", si lo hacemos por sintaxis son los comandos:

```
COMPUTE SPA=((100*0.1)+(75*2.0)+(50*14.9)+(25*32.0)+(0*51.0))/100.
COMPUTE IEP=((100*13.3)+(50*55.4)+(0*31.3))/100.
COMPUTE ICP=(SPA+IEP)/2.
```

Se generan 3 variables, que de hecho son constantes, con los valores de los índices: 17,05, 41,00 y 29,03.

2.1.2.4. *Recuento de valores*

Un procedimiento específico de cálculo consiste en contar para cada caso el número de veces que aparece algún valor o diversos valores en una lista de variables, numéricas o alfanuméricas. Se corresponde con el comando **COUNT** (menú Transformar / Contar valores dentro de los casos). Imaginemos que tenemos un listado de 15 bienes de consumo de equipamiento de los hogares, podríamos crear una variable que contara las veces que un hogar tiene cada bien (valor 1), la variable resultante podrá tener entre 0 (no tiene ningún bien) o 15 (los tiene todos).

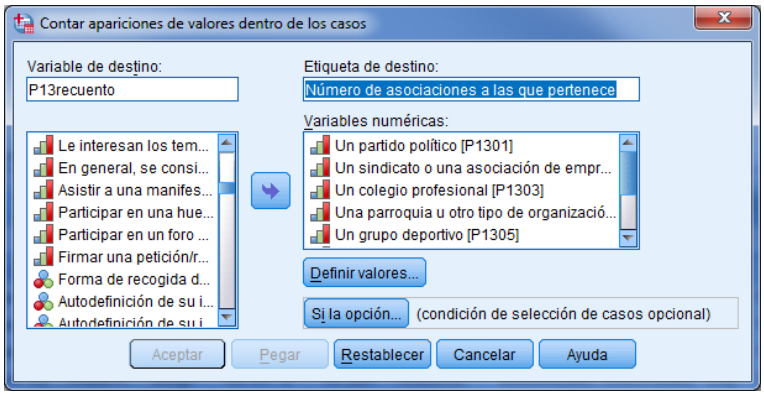
Con los datos de la matriz del CIS podemos considerar la pregunta 13 sobre participación en asociaciones

P.13 Las personas, algunas veces, pertenecen a ciertos grupos o asociaciones. Para cada uno de los que le voy a leer a continuación, dígame, por favor, si Ud.: ***(MOSTRAR TARJETA C)***.

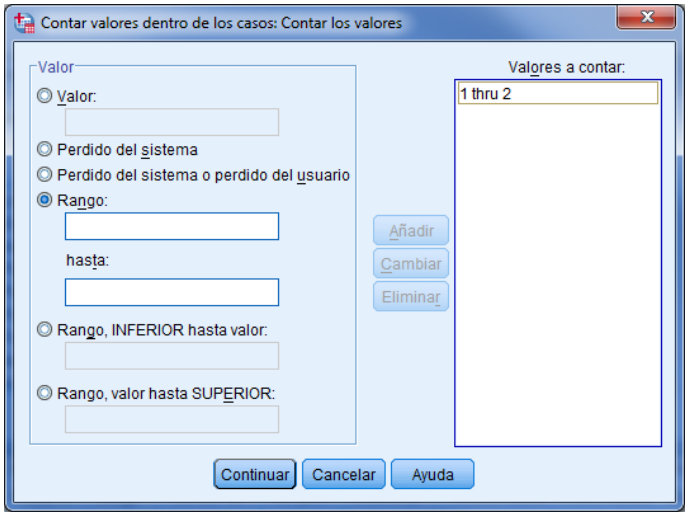
1. Pertenece y participa activamente
2. Pertenece, pero no participa activamente
3. Antes pertenecía, pero ahora no
4. Nunca ha pertenecido

	1	2	3	4	NC	
- Un partido político	1	2	3	4	9	(77)
- Un sindicato o una asociación de empresarios 1		2	3	4	9	(78)
- Un colegio profesional	1	2	3	4	9	(79)
- Una parroquia u otro tipo de organización/asociación religiosa	1	2	3	4	9	(80)
- Un grupo deportivo	1	2	3	4	9	(81)
- Un grupo cultural o de ocio	1	2	3	4	9	(82)
- Una organización de apoyo social o derechos humanos ..	1	2	3	4	9	(83)
- Una asociación juvenil o estudiantil	1	2	3	4	9	(84)
- Otro tipo de asociación voluntaria	1	2	3	4	9	(85)

Con las variables a las que da lugar la pregunta nos plantearnos como objetivo crear una variable sintética que cuente, para cada individuo, a cuantas asociaciones pertenece, es decir, si ha contestado 1 (pertenece y participa) o 2 (pertenece y no participa) a cada una de ellas. Como hay 9 preguntas la variable resultante tendrá valores entre 0 y 9. Entramos en el menú, seleccionamos las variables P1301 a P1309 y nombramos a la nueva variable P13recuento con la etiqueta Número de asociaciones a las que pertenece:



A continuación elegimos los valores de recuento en Definir valores y elegir el rango 1 hasta 2:



Clicamos en Continuar y Aceptar, y pedimos la tabla de frecuencias:

P13recuento Número de asociaciones a las que pertenece

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0	1558	62,8	62,8	62,8
	1	455	18,3	18,3	81,2
	2	223	9,0	9,0	90,2
	3	145	5,8	5,8	96,0
	4	53	2,1	2,1	98,1
	5	17	,7	,7	98,8
	6	19	,8	,8	99,6
	7	7	,3	,3	99,9
	8	2	,1	,1	100,0
	9	1	,0	,0	100,0
	Total	2480	100,0	100,0	

Se constata que la mayor parte de las personas no pertenece a ninguna asociación de las presentadas (62,8%) y muy pocas las que pertenecen a 4 o más.

► Ejercicio 10. Propuesto
A partir de la pregunta P10 sobre la frecuencia con que se discute de política obtener un índice de frecuencia calculando una variable con el recuento las veces que se responde 1 (A menudo) y 2 (Algunas veces) con relación a los tres grupos sociales.

2.1.2.5. *Transformaciones condicionales*

Para finalizar este recorrido por la transformación de las variables trabajaremos con un procedimiento de primera necesidad en el trabajo de análisis de la información cuantitativa: la creación de variables con transformaciones condicionales. Son situaciones donde se establecen determinadas condiciones en las características de las unidades y en función de su cumplimiento según una expresión lógica (verdadero o falso / perdido) asigna un valor a través de una expresión (dando el valor en concreto o ejecutando una fórmula de cálculo). La transformación condicional se puede utilizar en diversos comandos, pero nos detendremos sobre todo en el comando **IF** y en la estructura **DO IF ... END DIF**.

El comando **IF** que tiene la forma general siguiente:

IF [(]expresión lógica[)] variable de destino = expresión

donde los paréntesis de la expresión lógica aparece entre corchetes indicando que es optativo utilizarlos, aunque será obligatorio si la condición es compleja. El comando de hecho se parece al **COMPUTE** (Calcular variables) que vimos anteriormente. Prueba de ello es que el **IF** se obtiene a través del menú Transformar / Calcular variables / Si la opción.

A través de las transformaciones condicionales se construyen las variables tipológicas que combinan simultáneamente características de diversas variables (espacio de atributos) para definir diversos tipos. Es el caso de la construcción de la variable de clase social, del estilo de vida, de tipo de consumidor, etc.

Para ilustrar la utilización de ese procedimiento con el SPSS crearemos una variable (tipológica) de movilidad ocupacional intergeneracional a partir de relacionar el nivel ocupacional del padre con el alcanzado por el hijo/a. Las variables ocupacionales son respectivamente OCUPAPAD y OCUMAR11. Como paso previo pediremos la tabla de contingencia que cruza ambas variables (Analizar / Estadísticos descriptivos / Tablas cruzadas) para visualizar la información que se trabaja, ilustrar el procedimiento y luego poder verificar la creación de la nueva variable. Por convención en los análisis de movilidad social, en filas se coloca el origen social del padre y en columnas el del hijo/a. La tabla es la siguiente:

		OCUMAR11 Ocupación del hijo/a									
		1	2	3	4	5	6	7	8	9	Total
OCUPAPAD	1	13	19	13	3	12	0	2	1	6	69
Ocupación	2	4	75	19	4	18	0	10	5	2	137
del padre	3	10	34	58	13	46	1	8	15	10	195
	4	1	7	9	9	14	1	3	1	4	49
	5	18	34	36	15	98	6	26	11	28	272
	6	7	26	35	9	80	84	73	60	50	424
	7	12	44	64	15	121	9	121	48	70	504
	8	7	33	48	11	79	7	50	91	29	355
	9	2	12	8	5	25	7	24	20	53	156
Total		74	284	290	84	493	115	317	252	252	2161

1 Directores y gerentes; 2 Técnicos y profesionales científicos e intelectuales; 3 Técnicos; profesionales de apoyo; 4 Empleados contables, administrativos y otros empleados de oficina; 5 Trabajadores de los servicios de restauración, personales, protección y vendedores; 6 Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero; 7 Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción, excepto operadores de instalación; 8 Operadores de instalaciones y maquinaria, y montadores; 9 Ocupaciones elementales

La diagonal (en azul) define la inmovilidad o la reproducción social ocupacional donde el origen ocupacional del padre es el mismo que el del hijo/a. Los valores del triángulo inferior (en verde) corresponden a la movilidad ascendente, los hijos/as tienen un nivel ocupacional más alto que los padres. Finalmente el triángulo superior (en rojo) corresponde a la movilidad descendente, los hijos/as tiene menor nivel ocupacional.

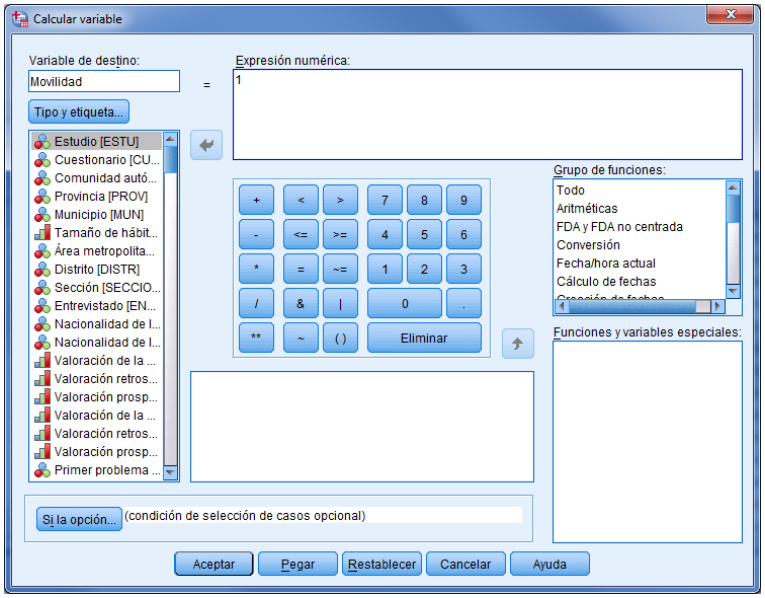
Para crear esta tipología de movilidad ocupacional utilizamos las transformaciones condicionales. En este caso establecemos 3 condiciones[29]:

- Si OCUPAPAD < OCUMAR11 entonces se da movilidad descendente (valor 1)
- Si OCUPAPAD = OCUMAR11 entonces se da inmovilidad (valor 2)
- Si OCUPAPAD > OCUMAR11 entonces se da movilidad ascendente (valor 3)

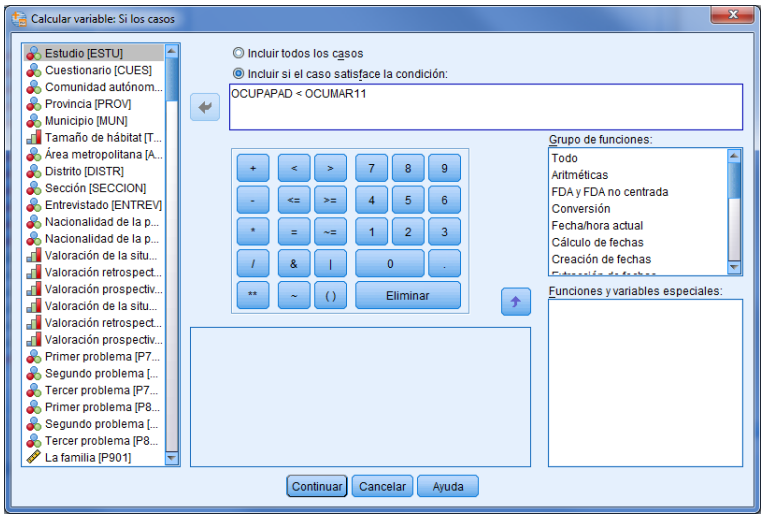
Todos los casos que no cumplan estas condiciones, es decir, los casos que corresponden valores perdidos de ambas variables, pasarán a ser valores perdidos del sistema. Para obtener la tabla anterior de 9 por 9 categorías se ha declarado valor perdido también al valor 10 (las Fuerzas Armadas).

Para traducir lo que comentamos en instrucciones para el SPSS podemos ir al menú Transformar / Calcular variables. En el cuadro de diálogo llamaremos a la nueva variable de destino Movilidad y pondremos como expresión numérica el 1.

[29] Como los valores van de 1, mayor nivel ocupacional, a 9, menor nivel, el sentido de la comparación es el inverso: un valor mayor entre origen y destino es movilidad descendente y un valor menor ascendente.



A continuación establecemos la condición que se ha de satisfacer para asignar el valor 1 a un individuo en la nueva variable (movilidad descendente), OCUPAPAD < OCUMAR11:



Para ejecutarlo primero presionamos Continuar y luego a Aceptar. Alternativamente podemos realizar esta tarea por sintaxis de la siguiente forma. En vez de clicar sobre Aceptar lo hacemos sobre Pegar. Nos engancha la instrucción siguiente en una ventana de sintaxis:

```
IF  (OCUPAPAD < OCUMAR11) Movilidad=1.
EXECUTE.
```

Como se puede comprobar, y con el tiempo y la experiencia con SPSS se verá más claramente, es más eficiente escribir esta instrucción directamente que realizar todo el recorrido anterior por el menú. Más aún si se tiene que repetir diversas veces para contemplar diversas situaciones que pueden ser muchas más de las tres que aquí estamos viendo. Adjuntada la primera instrucción la copiaremos dos veces más y las

modificaremos con las otras dos condiciones: inmovilidad, OCUPAPAD = OCUMAR11 y movilidad ascendente, OCUPAPAD > OCUMAR11:

```
IF  (OCUPAPAD < OCUMAR11) Movilidad=1.
IF  (OCUPAPAD = OCUMAR11) Movilidad=2.
IF  (OCUPAPAD > OCUMAR11) Movilidad=3.
EXECUTE.
```

Seleccionamos las cuatro líneas y las ejecutamos clicando sobre el icono de ejecución  o con las teclas <CTRL>+<R>. Se creará la nueva variable que tenemos que acabar de acondicionar con su diccionario. A continuación pedimos la tabla de frecuencias y se obtiene este resultado:

Movilidad Movilidad ocupacional intergeneracional

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Descendente	631	25,4	29,2	29,2
	2 Inmovilidad	602	24,3	27,9	57,1
	3 Ascendente	928	37,4	42,9	100,0
	Total	2161	87,1	100,0	
Perdidos	Sistema	319	12,9		
Total		2480	100,0		

Como se puede observar destaca la movilidad ocupacional absoluta ascendente (43%) como resultado del proceso de cambios que ha experimentado la sociedad española desde el periodo de industrialización a la fase postindustrial actual.

► Ejercicio 11. Propuesto
Realizar un análisis de la relación entre nivel educativo (variable ESTUDIOS) y la ocupación (variable OCUMAR11) de las personas entrevistadas. Proponer la creación de una tipología empírica que las relacione a partir de las frecuencias observadas en la tabla de contingencia.

► Ejercicio 12. Propuesto
Crear una variable tipológica que relacione el dinero y la felicidad, considerando las variables Escala de felicidad personal (P30) e Ingresos personales (P46). Para ello recodificar previamente cada una de las variables en tres categorías: feliz, ni feliz ni feliz, infeliz para la felicidad, y rico, ni rico ni pobre y pobre para los ingresos. Responder a la pregunta: ¿hasta qué punto el dinero hace la felicidad?

Nos podemos preguntar a continuación si estos resultados cambian cuando consideramos también a las madres, invisibilizadas en el ejercicio anterior, y en general en los análisis de movilidad social (Fachelli y López-Roldán, 2013, 2015). Para ello debemos resolver el tema de cómo determinar el "origen ocupacional de padres y madres". Una solución es aplicar el criterio de dominancia: se toma el mayor nivel ocupacional, el del padre o el de la madre. Crearemos en consecuencia una variable de ocupación dominante familiar con el nombre de OCUPAFAM.

Esta consideración implica realizar un ejercicio de análisis previo de homogamia ocupacional que podemos obtener cruzando la ocupación del padre y de la madre. Tal y como están definidos los valores perdidos de ambas variables, OCUPAPAD y

OCUPAMAD, dejaríamos de considerar muchos casos pues muchas madres solían estar clasificadas como "inactivas" en el pasado. Por otro lado el resto de valores que no precisan la ocupación en el caso del padre o en el caso de la madre se pueden recuperar si existe información de la ocupación de uno de los dos miembros. Para ello suprimiremos la declaración de valores perdidos y realizaremos el cruce con todos los valores de ambas variables:

		OCUPAMAD Ocupación de la madre a los 16 años de la persona entrevistada (CNO11)															
		1	2	3	4	5	6	7	8	9	10	95	96	97	98	99	Total
OCUPAPAD Ocupación del padre a los 16 años de la persona entrevistada (CNO11)	1	4	7	5	2	7	0	1	0	0	1	1	43	0	0	1	72
	2	0	41	13	4	6	1	2	0	2	0	0	67	2	0	0	138
	3	1	9	14	2	13	0	4	6	8	0	0	138	3	0	1	199
	4	0	3	3	1	8	0	0	0	1	0	0	33	0	1	0	50
	5	0	7	8	3	57	3	3	3	23	0	1	167	4	0	0	279
	6	0	5	1	1	13	78	5	7	8	1	0	301	8	1	0	429
	7	1	6	9	1	45	3	19	14	50	0	3	352	4	2	3	512
	8	0	2	6	1	23	3	7	21	25	0	1	261	5	0	3	358
	9	1	2	1	1	13	1	2	2	30	0	0	99	4	0	0	156
	10	0	1	0	1	0	0	1	1	1	0	0	14	0	0	0	19
	94	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	95	0	1	0	0	0	0	0	0	2	0	1	2	0	0	0	6
	96	0	0	3	0	6	2	0	1	6	0	0	27	0	0	0	45
	97	0	2	1	3	20	7	7	3	32	0	0	59	14	0	0	148
	98	1	1	1	1	4	1	0	1	4	0	0	27	1	1	1	44
	99	0	0	0	1	0	0	1	0	2	0	0	7	0	1	12	24
Total		8	87	65	22	216	99	52	59	194	2	7	1597	45	6	21	2480

1 Directores y gerentes; 2 Técnicos y profesionales científicos e intelectuales; 3 Técnicos; profesionales de apoyo; 4 Empleados contables, administrativos y otros empleados de oficina; 5 Trabajadores de los servicios de restauración, personales, protección y vendedores; 6 Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero; 7 Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción, excepto operadores de instalación; 8 Operadores de instalaciones y maquinaria, y montadores; 9 Ocupaciones elementales; 94 Sin ocupación, vivía de las rentas; 95 Parados; 96 Inactivos (ni ocupado, ni parado, o trabajo doméstico no remunerado, etc.); 97 No procede (no estaba presente, había fallecido, etc.); 98 N.S./No recuerda/Ocupación mal especificada; 99 N.C.

Se pueden identificar cuatro regiones en la tabla. En primer lugar, cuando existe información de la ocupación del padre y de la madre, de forma similar al ejemplo anterior de movilidad, definiremos la ocupación familiar así:

- Si OCUPAPAD < OCUPAMAD entonces OCUPAFAM la del padre.
- Si OCUPAPAD = OCUMAR11 entonces OCUPAFAM la del padre o la madre.
- Si OCUPAPAD > OCUMAR11 entonces OCUPAFAM la de la madre.

El resto de las regiones de la tabla definen estas situaciones:

- Si OCUPAPAD conocida y OCUPAMAD desconocida entonces OCUPAPAD.
- Si OCUPAPAD desonocida y OCUPAMAD conocida entonces OCUPAMAD.
- Si OCUPAPAD y OCUPAMAD desonocidas entonces sin datos.

¿Cómo traducirlo a SPSS? En primer lugar hemos comentado la existencia de 4 regiones o situaciones que tratar. Cada una de ellas se puede considerar por separado y en cada caso operar la condición de transformación necesaria para la creación de la variable de ocupación familiar. Existe un comando o estructura (**DO IF…END IF**) que ejecuta condicionalmente una o más transformaciones en subconjuntos de casos basados en expresiones lógicas. Su esquema es el siguiente:

```
DO IF [(]expresión lógica[)]
comandos de transformación
 [ELSE IF [(]expresión lógica [)]]
comandos de transformación
 [ELSE IF [(]expresión lógica [)]]
...
[ELSE]
comandos de transformación
END IF
```

DO IF establece una primera condición a partir de la cual se opera una transformación, optativamente se pueden establecer sucesivas condiciones con ELSE IF con sus correspondientes transformaciones. ELSE se puede utilizar dentro de la estructura para ejecutar transformaciones cuando las expresiones lógicas anteriores no se cumplen y así controlamos los casos no contemplados.

Apliquémoslo a nuestro caso. Solamente puede realizarse por sintaxis y sería la siguiente:

```
DO IF (OCUPAPAD <= 9 AND OCUPAMAD <= 9).
IF  (OCUPAPAD < OCUPAMAD) OCUPAFAM=OCUPAPAD.
IF  (OCUPAPAD = OCUPAMAD) OCUPAFAM=OCUPAPAD.
IF  (OCUPAPAD > OCUPAMAD) OCUPAFAM=OCUPAMAD.
ELSE IF (OCUPAPAD <= 9 AND OCUPAMAD >= 10).
COMPUTE OCUPAFAM=OCUPAPAD.
ELSE IF (OCUPAMAD <= 9 AND OCUPAPAD >= 10).
COMPUTE OCUPAFAM=OCUPAMAD.
ELSE.
COMPUTE OCUPAFAM=0.
END IF.
```

En la línea de DO IF se establece la primera condición (ocupación conocida de padre y madre) y en los 3 comandos IF siguientes se toma la decisión de qué ocupación se asigna a la nueva variable OCUPAFAM. Si no se conoce la ocupación de la madre, condición del primer ELSE IF, entonces se calcula que la ocupación de origen será la del padre. En el siguiente ELSE IF, de forma similar, si la ocupación del padre, no se conoce entonces se calcula que la ocupación de origen será la de la madre. Finalmente el resto de las situaciones con ELSE, es decir, no disponer de la ocupación del padre y de la madre, implicará que la nueva variable tenga el valor 0. Este valor además lo declararemos a continuación como valor perdido del usuario y deberemos completar el diccionario de la nueva variable con etiquetas, tipo y nivel de medición. La tabla de frecuencias será:

OCUPAFAM Ocupación dominante de los padres y las madres

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Directores/as y gerentes	76	3,1	3,3	3,3
	2 Técnicos/as y profesionales científicos/as e intelectuales	177	7,1	7,7	10,9
	3 Técnicos/as; profesionales de apoyo	222	9,0	9,6	20,5
	4 Empleados/as contables, administrativos/as y otros empleados/as de oficina	57	2,3	2,5	23,0
	5 Trabajadores/as de los servicios de restauración, personales, protección y vendedores/as	386	15,6	16,7	39,7
	6 Trabajadores/as cualificados/as en el sector agrícola, ganadero, forestal y pesquero	426	17,2	18,4	58,1
	7 Artesanos/as y trabajadores/as cualificados/as de las industrias manufactureras y la construcción, excepto operadores/as	465	18,8	20,1	78,2
	8 Operadores/as de instalaciones y maquinaria, y montadores/as	324	13,1	14,0	92,2
	9 Ocupaciones elementales	180	7,3	7,8	100,0
	Total	2313	93,3	100,0	
Perdidos	0 Sin datos	167	6,7		
Total		2480	100,0		

Queda analizar la movilidad absoluta intergeneracional y construir como antes la variable de movilidad (Movilidad2), ahora entre el origen ocupacional de los padres y las madres y el destino de los hijos y las hijas. La tabla de movilidad es:

		OCUMAR11 Ocupación del hijo/a									
		1	2	3	4	5	6	7	8	9	Total
OCUPAFAM Ocupación dominante de los padres y las madres	1	14	21	13	3	13	0	2	1	6	73
	2	6	88	30	5	23	1	12	6	4	175
	3	13	40	61	15	50	3	8	17	12	219
	4	1	6	10	10	16	2	3	3	5	56
	5	18	41	40	19	129	8	48	24	48	375
	6	7	21	37	7	80	84	75	61	49	421
	7	11	38	60	14	106	11	111	45	62	458
	8	6	30	43	11	73	7	40	83	28	321
	9	2	10	12	5	29	6	30	21	64	179
Total		78	295	306	89	519	122	329	261	278	2277

y las instrucciones son:

```
IF  (OCUPAFAM < OCUMAR11) Movilidad2=1.
IF  (OCUPAFAM = OCUMAR11) Movilidad2=2.
IF  (OCUPAFAM > OCUMAR11) Movilidad2=3.
```

que completamos con el diccionario de los datos y sacando la tabla de frecuencias:

Movilidad2 Movilidad ocupacional intergeneracional

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Descendente	722	29,1	31,7	31,7
	2 Inmovilidad	644	26,0	28,3	60,0
	3 Ascendente	911	36,7	40,0	100,0
	Total	2277	91,8	100,0	
Perdidos	Sistema	203	8,2		
Total		2480	100,0		

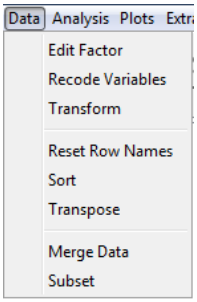
Como resultado observamos que la movilidad ascendente disminuyó algo, pasando del 43 al 40%, debido al hecho que el criterio de dominancia tiende a elevar la posición de origen al elegir la más alta entre el padre y la madre; y como las posiciones de origen son más altas las posibilidades de ascender socialmente serán menores.

Como hemos ido viendo a lo largo de este apartado, la realización de transformaciones con las variables implica modificar o crear otras nuevas que van ampliando nuestro fichero de datos como pusimos de manifiesto al inicio de este capítulo al hablar del proceso de datos. Ello implica gestionar cómo guardar estos datos. Una práctica recomendable es mantener una copia de la fuente de datos original y crear la matriz ampliada guardándola con otro nombre. En nuestro caso todas las variables que hemos ido generando se encuentran en la matriz CIS3041+.sav.

Conviene observar también que los datos generados se han obtenido en general desde el menú en una dinámica de trabajo interactiva lo que puede representar una limitación de cara a replicar el trabajo realizado. Para volver a realizar los ejercicios vistos disponemos del propio manual, pero en la práctica de la investigación, revisar o rehacer la generación de los datos y su análisis requiere registrarlo. Una forma de hacerlo es guardar sistemáticamente los archivos de resultados que contienen la sintaxis y los resultados de su ejecución. Pero volver a ejecutarlos por el menú para traducir aquellos comandos y resultados puede resultar complicado, largo y laborioso. La alternativa es guardar archivos de sintaxis con todas las tareas realizadas que al ser ejecutados de nuevo, en cuestión de segundos, generan todo el trabajo de horas que representó cuando se diseñaron originalmente. Así hemos trabajado nosotros y hemos guardado todas las transformaciones que se han visto en el capítulo en el programa de sintaxis Transformar.sps que se puede consultar en la página web de este capítulo.

2.2. Transformación de los datos con R

Comentaremos los distintos procedimientos que se presentan en el menús de Deducer: Data, que incluye algunos procedimientos destinados al tratamiento de ficheros, ya sea en su interior ya sea para combinarlo con otros, y de transformación para la creación de variables.



2.2.1. *Tratamiento de ficheros con R*

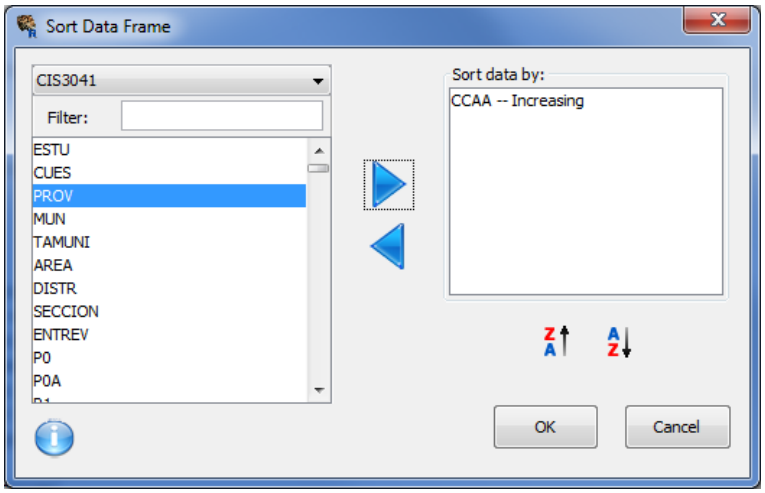
Distinguiremos dos tipos de procedimientos de gestión y transformación de archivos, los destinados al tratamiento de datos en el interior de un fichero y al tratamiento de datos entre ficheros que se relacionan.

2.2.1.1. *Tratamiento de datos en el interior de un fichero*

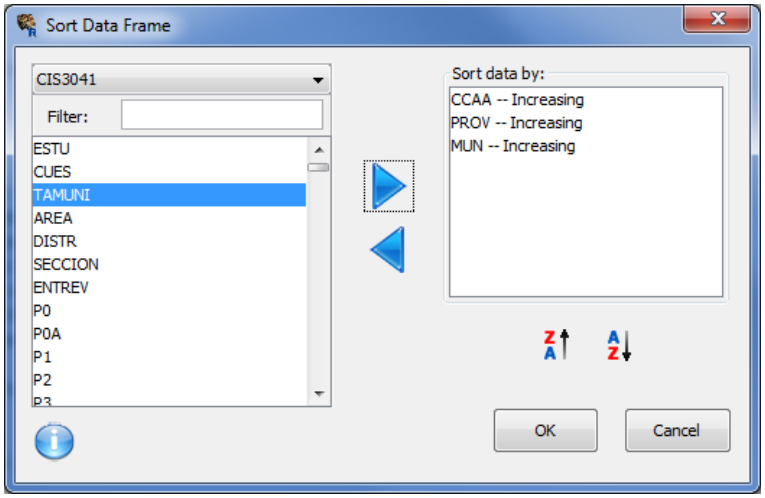
Ordenar casos

El comando de ordenar casos (menú Data / Sort) permite la reordenación de los casos del fichero activo según los valores especificados en una o más variables, numéricas o alfanuméricas (cadena, para éstas el orden es el alfabético). Los casos pueden ser reordenados en orden ascendente, por defecto, o descendente.

Con la matriz de datos CIS3041.rda vemos que los casos están inicialmente ordenados según el número del cuestionario (variable CUES). Como ejercicio podemos ordenar el archivo según el lugar de la entrevista. Un primer criterio sería por ejemplo ordenar el archivo según la Comunidad Autónoma (variable CCAA) en orden ascendente:



Obsérvense los cambios en el archivo de datos. Si queremos precisar más podemos poner además de la variable CCAA, la variable de la provincia (PROV) y del municipio (MUN), todas en orden ascendente. Las introduciremos por este orden:

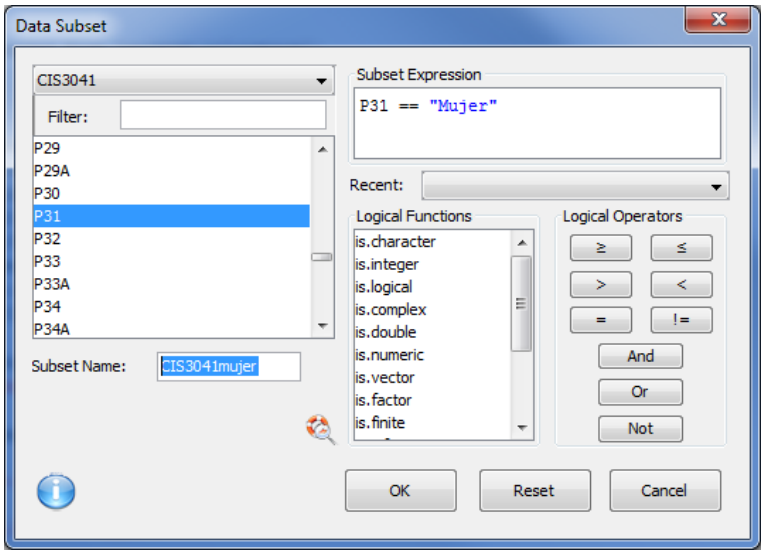


La ordenación de un archivo de pequeñas dimensiones es instantánea pero con archivos de millones de registros puede tardar minutos, en este sentido es muy útil tener la base de datos ordenada según un criterio si se utiliza de forma habitual.

Veremos también que la ordenación de un archivo es un paso previo necesario en diversos procedimientos de tratamiento de datos.

Seleccionar casos

A menudo, cuando trabajamos con una base de datos nos interesa obtener información sobre los individuos que satisfacen determinadas condiciones. Nos puede interesar, por ejemplo, estudiar diversas variables pero sólo para los individuos con determinadas características: de sexo femenino, los que piensan votar, los que tienen un bajo nivel de ingresos, etc. Con Deducer podemos seleccionar el subconjunto de los individuos que satisfacen una determinada condición de forma que se crea un objeto nuevo con los datos seleccionados, un nuevo *dataframe*. Como ejercicio podemos seleccionar los casos de las personas entrevistadas que son mujeres. En el cuadro de diálogo del menú Data / Subset seleccionamos variable del sexo (la P31) y la pasamos a la derecha haciendo doble-clic. Para seleccionar a las mujeres escribiremos con el teclado o con los botones del cuadro de diálogo: =="Mujer"[30]:



Construida la condición podemos cambiar el nombre (Subset Name) que por defecto se asignará al objeto con los datos de la selección, por ejemplo CIS3041mujer. Clicaremos sobre OK y se ejecutará, es decir, dispondremos en el espacio de trabajo de una nueva matriz con la información de los casos que corresponden a las mujeres y que podemos visualizar desde el visor de datos. Si queremos obtener por ejemplo una tabla de frecuencias de una variable en el cuadro de diálogo de Frequencies podemos elegir en cada momento la matriz con la que queremos trabajar, si con toda la muestra (CIS3041) o con esta submuestra de mujeres que acabamos de crear (CIS3041mujer).

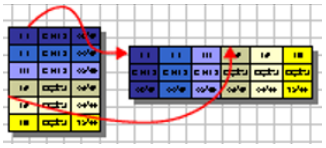
Transponer

La transposición de una matriz implica convertir los casos (las filas) en variables, y las variables (las columnas) en casos. Al hacerlo se crea un nuevo archivo de datos y automáticamente los nombres de las variables y los nombres de las filas.

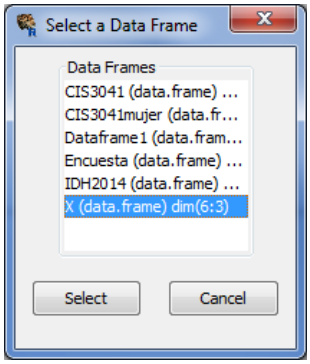
[30] El igual en R es un doble signo de igual.

Para ilustrar este procedimiento y los que vienen a continuación, trabajaremos con unas pequeñas matrices de datos que permitirán ver mejor cada una de las tareas. Consideraremos la matriz de datos X.rda que contiene la situación laboral de 6 individuos asalariados en relación a 2 variables de sus condiciones de empleo: Contrato y Salario. Se puede abrir directamente desde el editor de datos de Deducer:

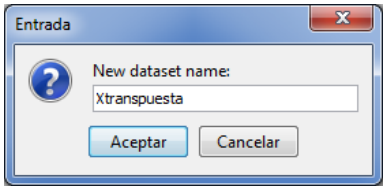
	ID	Contrato	Salario
1	1	Fijo	1200
2	2	Temporal	1000
3	3	Fijo	3000
4	4	Temporal	1000
5	5	Fijo	1200
6	6	Fijo	1500



Para transponerla iremos al menú Data / Transpose, nos pedirá elegir la matriz de datos:



Una vez seleccionada nos pediré darle un nombre a la nueva matriz de datos que se creará, por ejemplo Xtranspuesta:



Para ver el resultado volvemos al editor de datos y buscamos la nueva matriz:

	V1	V2	V3	V4	V5	V6
ID	1	2	3	4	5	6
Contrato	Fijo	Temporal	Fijo	Temporal	Fijo	Fijo
Salario	1200	1000	3000	1000	1200	1500

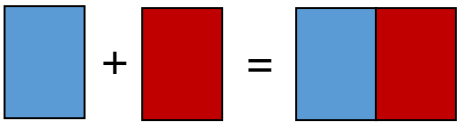
2.2.1.2. *Tratamiento de datos entre ficheros que se relacionan*

Veremos a continuación otras tareas de manipulación de matrices de datos que implican relacionar dos o más archivos: la fusión.

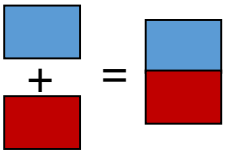
Fusionar archivos

La fusión o unión de archivos da lugar a dos alternativas:

- Añadir variables. Se fusiona el archivo de datos activo con otro que contiene los mismos casos pero variables diferentes.



- Añadir casos. Se fusiona el archivo de datos activo con otro que contiene las mismas variables pero casos diferentes.



Realizaremos un pequeño ejercicio con la matriz Y.rda que contine 6 casos y 4 variables, Edad y Sexo son características individuales sociodemográficas y Sector y Tamaño hacen referencia a características laborales de la empresa:

	ID	Edad	Sexo	Sector	Tamaño
1	1	23	Mujer	Servicios	20
2	2	35	Varón	Primario	1
3	3	48	Varón	Industria	100
4	4	55	Mujer	Industria	500
5	5	28	Varón	Construcción	50
6	6	20	Varón	Servicios	5

Para el ejercicio de unir variables consideraremos dos matrices iniciales separadas con la información sociodemográfica (YA.rda) y la información de la empresa (YB.rda). Para el ejercicio de unir casos disponemos de dos matrices separadas con los tres primeros casos (Y1.rda) y los tres últimos (Y4-6.rda). Las abrimos desde Deducer.

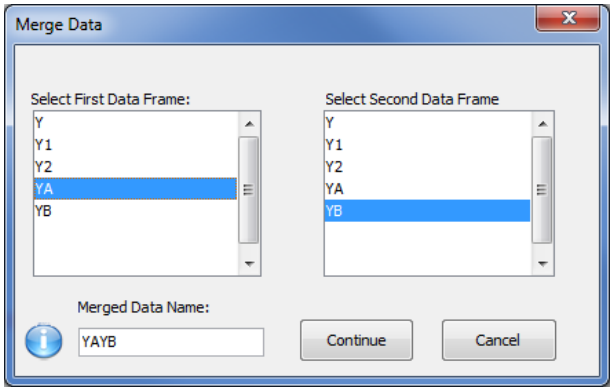
	ID	Edad	Sexo	Sector	Tamaño
1	1	23	Mujer	Servicios	20
2	2	35	Varón	Primario	1
3	3	48	Varón	Industria	100
4	4	55	Mujer	Industria	500
5	5	28	Varón	Construcción	50
6	6	20	Varón	Servicios	5

YA YB

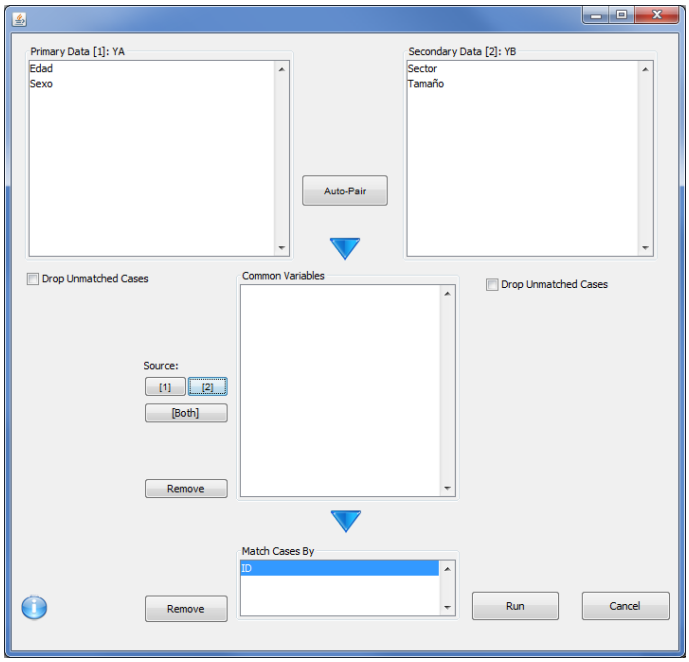
	ID	Edad	Sexo	Sector	Tamaño
1	1	23	Mujer	Servicios	20
2	2	35	Varón	Primario	1
3	3	48	Varón	Industria	100
4	4	55	Mujer	Industria	500
5	5	28	Varón	Construcción	50
6	6	20	Varón	Servicios	5

Y1 Y2

La fusión se realiza a través del menú Data / Merge. Se abre el cuadro de diálogo donde aparecen las matrices del espacio de trabajo que previamente habremos cargado:

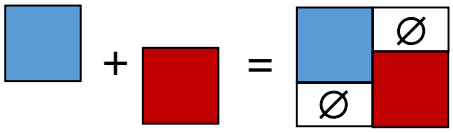


Primero realizaremos la fusión de YA con YB tarea que implica añadir las variables de YB a las existentes en YA. A la nueva matriz le llamamos YAYB. Clicamos sobre continuar y nos aparece el cuadro de diálogo de la fusión:



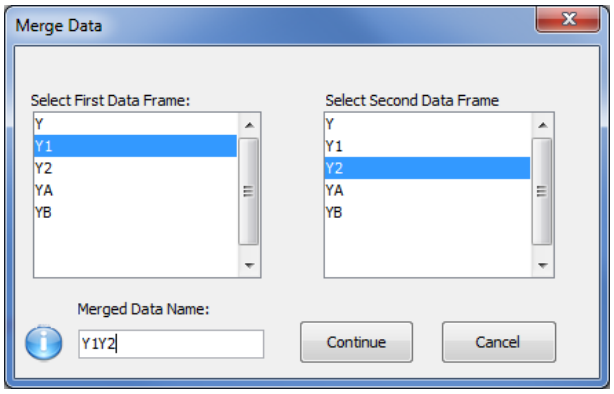
Vemos tres recuadros con las variables propias de cada archivo y las que son comunes. En este último estaba la variable ID que utilizamos como variable de control del emparejamiento de los casos. Pasa fusionar siempre es conveniente disponer de una variable clave que identifique a cada unidad en cada uno de los archivos a unir, de esta forma se irá emparejando la información a partir del control de la coincidencia del mismo caso. En nuestro ejemplo este papel lo juega la variable ID y se coloca en el recuadro Match Cases By: después de elegir si la variable es la del primer archivo: [1], del segundo: [2], o de ambos [b] y en este caso creará dos versiones de la variable. Una vez ejecutado con Run tendremos como resultado la misma información de la matriz Y.

Conviene tener presente que todos los casos desemparejados, es decir, los que están en una matriz y no en la otra, sea la que sea, tendrán valores perdidos en la fusión para las variables donde no tienen información, serán vacíos (∅) en la nueva matriz:

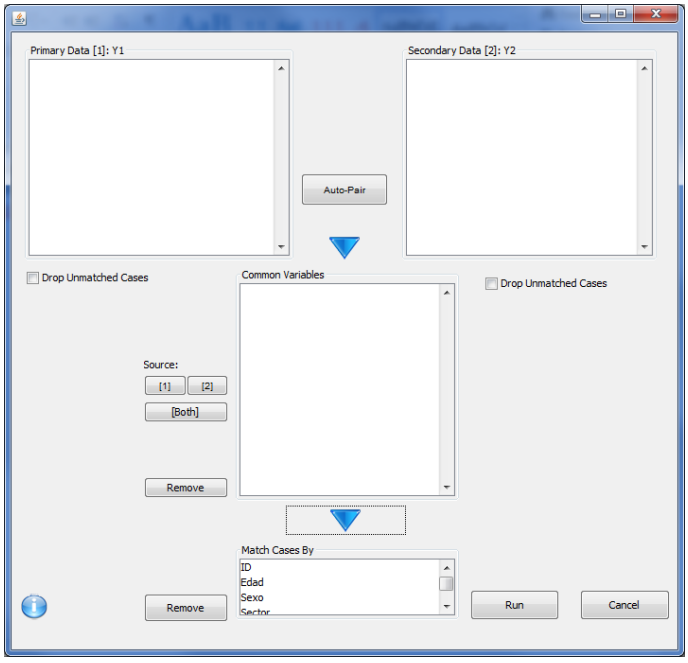


Con Deducer existe la posibilidad de eliminar a priori los casos que no se emparejan (Drop Unmatched Cases). Si dos variables representan el mismo elemento, pero se denominan de manera diferente en las dos matrices de datos, se pueden combinar mediante la selección de las dos variables y haciendo clic en la flecha hacia abajo y ubicarlas unidas en el recuadro de Common Variables.

Realizaremos ahora el segundo caso de fusión, el de añadir casos. Elegiremos la matriz Y1 que contiene los 3 primeros casos y le añadiremos Y2 con los 3 últimos. A la nueva matriz le llamamos Y1Y2:



En este caso todas las variables son comunes. Las variables que quedan desemparejadas, porque están en un fichero y no en el otro, no se incluirán en el archivo fusionado. Debemos pasar todas las variables del recuadro Common Variables a Match Cases By clicando sobre la flecha hacia abajo:



De nuevo ejecutando el procedimiento reproducimos la matriz original Y.

2.2.2. *Transformación de variables*

Después de ver distintas operaciones de tratamiento de una matriz en su conjunto nos centramos en aquellas tareas de transformación donde se implican variables concretas de la matriz, de forma individual o relacionándolas con otras. Son diversos los comandos destinados a la transformación de las variables existentes, bien sea para su modificación o bien por la generación o creación de nuevas variables. La construcción de tipologías y de índices a partir de diversas variables será una de las necesidades

frecuentes del análisis, la recodificación de los valores de las variables para agrupar valores o reducir la escala de medida es otra tarea inmediata que conlleva el análisis.

En todo ejercicio de creación de variables hay que tener presente el comportamiento de los valores perdidos en dos momentos: antes y después de crear las variables. Antes, hay que tener en cuenta que si las variables contienen valores perdidos, en las nuevas variables éstos aparecerán como valores perdidos si no se tratan específicamente. Por otra parte, cuando creamos una variable nueva debemos prever y controlar la generación no deseada de valores perdidos como resultado de una operación en la que las transformaciones no se aplican de hecho en todos los casos que inicialmente queremos considerar. Si alguna transformación no se aplica a un caso concreto el valor en la variable creada que aparecerá será un valor perdido.

Hay que tener presente finalmente que toda generación de variables requiere a menudo completar su diccionario (tipo de variable y ordenación de categorías).

2.2.2.1. *Recodificación de variables*

La recodificación de variables permite cambiar los valores actuales de las variables por otros nuevos. La recodificación puede significar estrictamente un cambio de uno o más valores por otros, o bien la combinación o la agrupación de rangos de valores en nuevas categorías.

Por otro lado la recodificación se puede realizar optando por mantener la variable original y generando una nueva con otra nombre que tendrá los valores recodificados, o bien optando por sustituir la variable que se está recodificando por la nueva variable con los nuevos criterios de codificación y con el mismo nombre de variable.

Consideraremos la matriz de datos CIS3041 y realizaremos dos ejercicios de recodificación: a partir de una variable cualitativa y a partir de una cuantitativa.

El primer paso para realizar una recodificación es definir los criterios de recodificación y observar los valores de las variables extrayendo la tabla de frecuencias. Consideramos en primer lugar la variable OCUMAR11, la categoría ocupacional de la persona entrevistada según la CNO de 2011 (Clasificación Nacional de Ocupaciones)[31] . Su tabla de frecuencias aparece a continuación. Las etiquetas abreviadas de la variable se corresponden a las descripciones siguientes:

Director: Directores y gerentes; Técnico: Técnicos y profesionales científicos e intelectuales; Apoyo: Técnicos; profesionales de apoyo; Administrativos: Empleados contables, administrativos y otros empleados de oficina; Servicios: Trabajadores de los servicios de restauración, personales, protección y vendedores; Cualificados agrícolas: Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero; Cualificados industria: Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción, excepto operadores de instalación; Operadores: Operadores de instalaciones y maquinaria, y montadores; Elementales: Ocupaciones elementales; NA: Sin ocupación, vivía de las rentas; Parados; Inactivos (ni ocupado, ni parado, o trabajo doméstico no remunerado, etc.); No procede (no estaba presente, había fallecido, etc.); N.S./No recuerda/Ocupación mal especificada; N.C.

[31] La CNO (http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft40%2Fcno11%2F&file=inebase&L=0) es la adaptación española de la clasificación internacional ISCO (*International Standard Classification of Occupations*) de la OIT (http://www.ilo.org/public/spanish/bureau/stat/isco/), o CIUO, que tiene varios niveles de desagregación, hasta 5 y se codifica a 4 dígitos. Aquí se presenta con un 1 solo dígito. La variable P40 de la matriz CIS3041.sav es la CNO 2011 a tres dígitos. Por tanto, de hecho OCUMAR11 es ya una variable que ha sido recodificada (agrupada) a un solo dígito.

Frequencies (OCUMAR11)

	Value	# of Cases	%	Cumulative %
1	Director	84	3.40	3.40
2	Tecnico	309	12.60	16.10
3	Apoyo	325	13.30	29.40
4	Administrativo	100	4.10	33.40
5	Servicios	559	22.90	56.30
6	Cualificado agricola	132	5.40	61.70
7	Cualificado industria	359	14.70	76.40
8	Operadores	274	11.20	87.60
9	Elemental	294	12.00	99.60
10	Militar	10	0.40	100.00

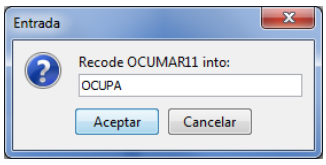
Case Summary (OCUMAR11)

	Valid	Missing	Total	% Missing
1	2446.00	34.00	2480.00	1.40

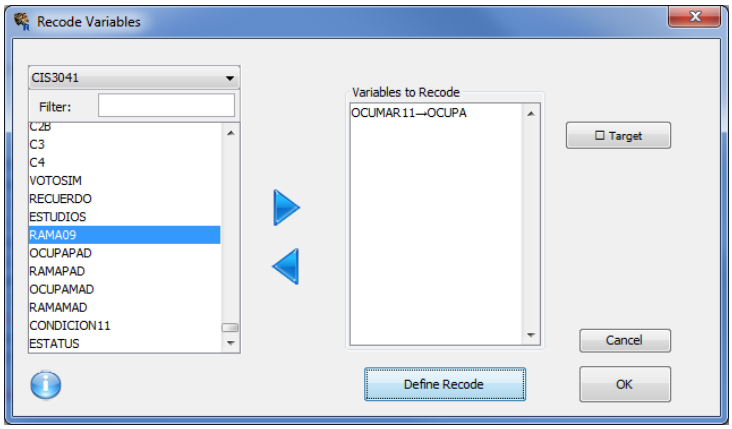
El objetivo es disponer de una variable ocupacional con un número más reducido que las 10 categorías de la variable original. Consideraremos una agrupación en 4 categorías ocupacionales más una categoría de valores perdidos según los siguientes criterios:

1. Clase alta o categoría ocupacional alta: Director y Técnico.
2. Clase media o categoría ocupacional intermedia: Apoyo y Administrativo.
3. Trabajadores cualificados o categoría media baja: Servicios, Cualificado agrícola y Cualificado industria.
4. Trabajadores no cualificados o categoría baja: Operadores y Elemental.
5. Valores perdidos: Militar (que se unirán a los 34 casos existentes).

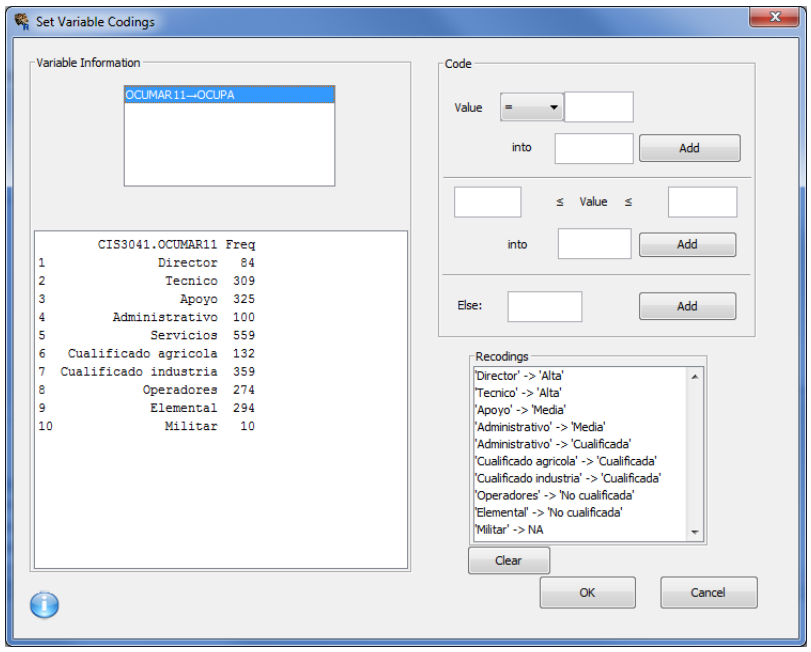
Entramos en el menú del procedimiento Data / Recode Variables elegimos la variable OCUMAR11 para pasarla al recuadro de la derecha de Variables to Recode. Automáticamente le asigna el mismo nombre indicando que recodificará en la misma variable. En general, si no se tiene la certeza para actuar de esta manera, preferiremos crear una nueva variable. Para ello seleccionamos la línea y clicamos sobre Target para cambiar el nombre de destino de la variable, escribimos el nombre de la nueva variable, por ejemplo OCUPA y clicamos sobre Aceptar:



El cuadro de diálogo inicial aparece de esta forma:



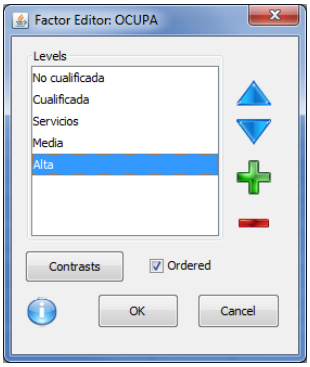
A continuación debemos especificar los criterios de recodificación en Define Recode:



Los criterios de recodificación que hemos comentado se trasladan de la forma siguiente. Primeramente clicaremos sobre la pareja de variables que aparece en el recuadro Variable Information veremos que para las variables numéricas se muestra una tabla de percentiles y para las variables cualitativas, como es el caso, una tabla de frecuencias. Con variables tipo factor no podemos utilizar el rango entre valores, deberemos escribir cada valor exactamente (copiaremos el texto que tenemos a la izquierda) y especificaremos el nuevo valor, el nuevo texto:

- En el primer caso sería escribir:
  Value = Director into Alta y clicar Add
  Value = Tecnico into Alta y clicar Add.
  Así definimos que Directores y gerentes junto a Tecnicos y profesionales, se unan en una sola categoría de clase ocupacional alta, codificada con Alta en la nueva.
- Lo mismo repetimos para los otros tres grupos ocupacionales: Media, Cualificado y No cualificado.
- En el último caso: Value = Militares into NA y clicar Add.
  El último valor lo consideraremos como valor perdido junto con los valores perdidos que ya tiene la variable identificados con el símbolo NA en la matriz.

Le damos a OK en esta ventana y de nuevo en la siguiente para ejecutar la recodificación.



Para ver el resultado necesitamos pedir la tabla de frecuencias, pero previamente es preciso mejorar el diccionario de los datos ordenando las etiquetas, y eliminando la Militar que aparece con frecuencia cero, y marcando su carácter ordinal.

El resultado final es el siguiente:

Frequencies (OCUPA)

	Value	# of Cases	%	Cumulative %
1	Alta	393	16.10	16.10
2	Media	325	13.30	29.50
3	Servicios	559	22.90	52.40
4	Cualificada	591	24.30	76.70
5	No cualificada	568	23.30	100.00

Case Summary (OCUPA)

	Valid	Missing	Total	% Missing
1	2436.00	44.00	2480.00	1.80

► Ejercicio 13. Propuesto

El INE en su informe *Introducción a la CNO-11* comenta el concepto de ocupación (http://www.ine.es/daco/daco42/clasificaciones/Introduccion_CNO11.V02.pdf) y distingue entre empleo y competencias como dos dimensiones fundamentales que lo estructuran. Las competencias distinguen a su vez dos dimensiones: la especialización y el nivel de competencias. Éste último tiene 4 grados (asociados teóricamente a los niveles educativos formales) que se corresponden con las categorías ocupacionales a 1 dígito de la forma siguiente:

Cuadro 1: Correspondencia entre los Grandes Grupos de la CIUO-08 y el nivel de competencias

Grandes Grupos CIUO-08	Nivel de competencias
1 - Directores y gerentes	3, 4
2 - Profesionales científicos e intelectuales	4
3 - Técnicos y profesionales de nivel medio	3
4 - Personal de apoyo administrativo 5 - Trabajadores de los servicios y vendedores de comercios y mercados 6 - Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros 7 - Oficiales, operarios y artesanos de artes mecánicas y de otros oficios 8 - Operadores de instalaciones y máquinas y ensambladores	2
9- Ocupaciones elementales	1
0- Ocupaciones militares	1, 2, 4

Fuente: INE

De acuerdo con esta tabla, sin considerar el 0 de las ocupaciones militares y asignando a los directores y gerentes sólo el nivel 4, agrupar los grandes grupos ocupacionales (variable OCUMAR11 de la matriz CIS3041.sav) en los 4 niveles de competencias. Completar igualmente el diccionario de las variables y extraer la tabla de frecuencias para comprobar el resultado.

Un segundo ejemplo de recodificación tendrá en cuenta una variable cuantitativa, la edad (variable P32). Es habitual trabajar con la edad agrupada en intervalos de 5 o 10 años, o en grandes grupos de edad (jóvenes, adultos, mayores). Así la variable original cuantitativa reduce su escala y permite trabajarla con menos categorías como una

variable cualitativa ordinal. Se propone crear una nueva variable de edad (Edad10) con una agrupación en intervalos según estos criterios:

1. 18 a 24
2. 25 a 34
3. 35 a 44
4. 45 a 54
5. 55 a 64
6. 65 y más

La variable original no tiene valores perdidos. La tabla de distribución de frecuencias original es la siguiente:

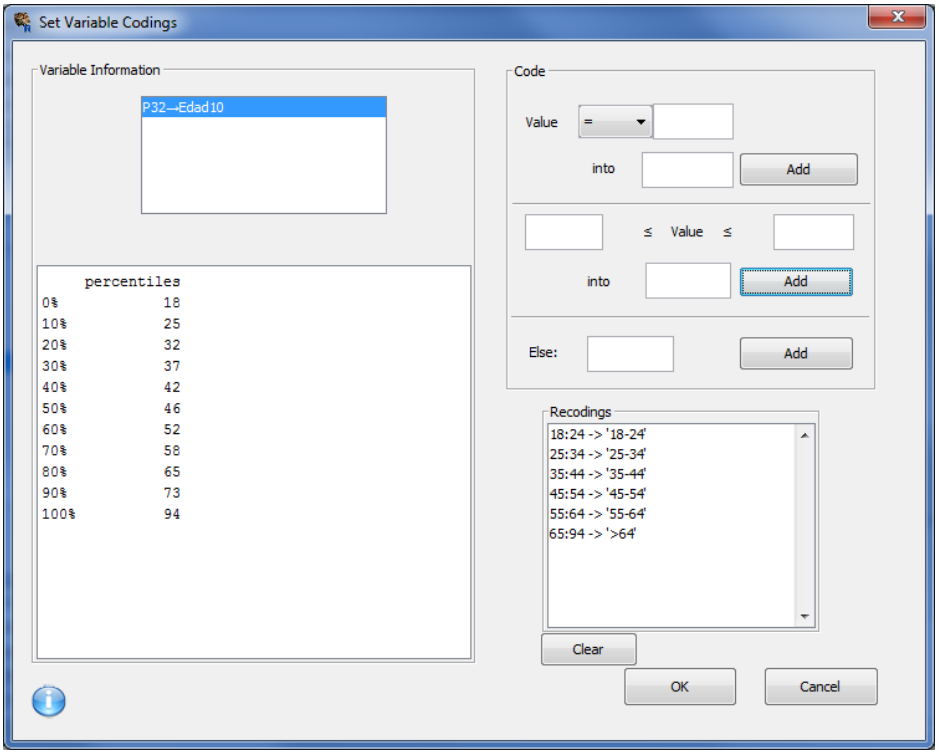
Frequencies (P32)

	Value	# of Cases	%	Cumulative %
1	18	32	1.30	1.30
2	19	32	1.30	2.60
3	20	18	0.70	3.30
4	21	28	1.10	4.40
5	22	38	1.50	6.00
6	23	25	1.00	7.00
7	24	27	1.10	8.10
8	25	51	2.10	10.10
9	26	42	1.70	11.80
10	27	40	1.60	13.40
11	28	23	0.90	14.40
12	29	39	1.60	15.90
13	30	46	1.90	17.80
14	31	48	1.90	19.70
15	32	41	1.70	21.40
16	33	47	1.90	23.30
17	34	53	2.10	25.40
18	35	51	2.10	27.50
19	36	37	1.50	29.00
20	37	48	1.90	30.90
21	38	47	1.90	32.80
22	39	46	1.90	34.60
23	40	48	1.90	36.60
24	41	43	1.70	38.30
25	42	57	2.30	40.60
26	43	61	2.50	43.10
27	44	71	2.90	45.90
28	45	51	2.10	48.00
29	46	51	2.10	50.00
30	47	45	1.80	51.90
31	48	42	1.70	53.50
32	49	45	1.80	55.40
33	50	57	2.30	57.70
34	51	33	1.30	59.00
35	52	34	1.40	60.40
36	53	49	2.00	62.30
37	54	56	2.30	64.60
38	55	34	1.40	66.00
39	56	36	1.50	67.40
40	57	42	1.70	69.10
41	58	29	1.20	70.30
42	59	27	1.10	71.40
43	60	48	1.90	73.30
44	61	27	1.10	74.40
45	62	33	1.30	75.70
46	63	39	1.60	77.30
47	64	40	1.60	78.90
48	65	48	1.90	80.80
49	66	37	1.50	82.30
50	67	39	1.60	83.90
51	68	24	1.00	84.90
52	69	31	1.20	86.10
53	70	36	1.50	87.60
54	71	27	1.10	88.70
55	72	28	1.10	89.80
56	73	18	0.70	90.50
57	74	21	0.80	91.40
58	75	19	0.80	92.10
59	76	20	0.80	92.90
60	77	18	0.70	93.70
61	78	25	1.00	94.70
62	79	16	0.60	95.30
63	80	17	0.70	96.00
64	81	17	0.70	96.70
65	82	14	0.60	97.30
66	83	15	0.60	97.90
67	84	13	0.50	98.40
68	85	11	0.40	98.80
69	86	8	0.30	99.20
70	87	4	0.20	99.30
71	88	5	0.20	99.50
72	89	4	0.20	99.70
73	90	2	0.10	99.80
74	91	2	0.10	99.80
75	92	1	0.00	99.90
76	94	3	0.10	100.00

Case Summary (P32)

	Valid	Missing	Total	% Missing
1	2480.00	0.00	2480.00	0.00

Siguiendo el protocolo que vimos anteriormente especificaremos en particular los criterios de recodificación ahora pudiendo utilizar el rango de valores:



La tabla de frecuencias resultante después de completar el diccionario de los datos: cambiar de *character* a *factor* y ordenar los valores, es la siguiente:

Frequencies (Edad10)

	Value	# of Cases	%	Cumulative %
1	18-24	200	8.10	8.10
2	25-34	430	17.30	25.40
3	35-44	509	20.50	45.90
4	45-54	463	18.70	64.60
5	55-64	355	14.30	78.90
6	>64	523	21.10	100.00

Case Summary (Edad10)

	Valid	Missing	Total	% Missing
1	2480.00	0.00	2480.00	0.00

► Ejercicio 14. Propuesto

Recodificar la variable P15 de autoposicionamiento ideológico en tres categorías que agrupen los valores 1 a 3, 4 a 6 i 7 a 10.

Si con los datos de la encuesta del CIS nos preguntamos ¿cuáles son los ingresos medios de los hogares de los entrevistados? Para responder a esta pregunta deberíamos tener la variable de ingresos como cuantitativa y en la encuesta se pregunta por intervalos de forma cualitativa. Una alternativa es calcular la media a partir de la marca de clase de cada intervalo para lo que deberemos recodificar la variable. La distribución de la variable de ingresos (P45) es la siguiente:

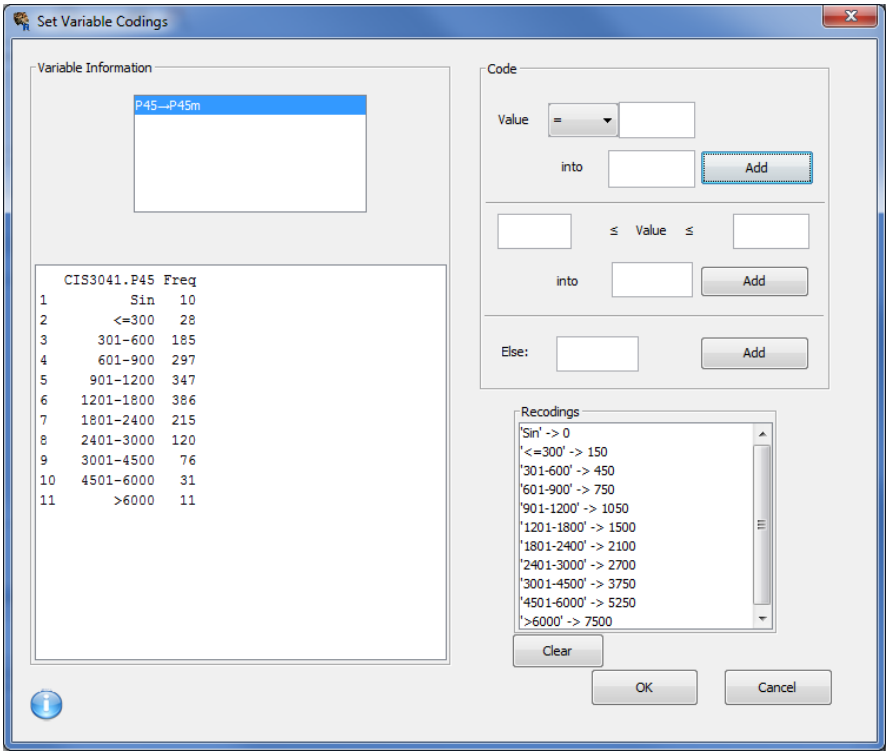
Frequencies (P45)

	Value	# of Cases	%	Cumulative %
1	Sin	10	0.60	0.60
2	=300	28	1.60	2.20
3	301-600	185	10.80	13.10
4	601-900	297	17.40	30.50
5	901-1200	347	20.30	50.80
6	1201-1800	386	22.60	73.40
7	1801-2400	215	12.60	86.00
8	2401-3000	120	7.00	93.10
9	3001-4500	76	4.50	97.50
10	4501-6000	31	1.80	99.40
11	>6000	11	0.60	100.00

Case Summary (P45)

	Valid	Missing	Total	% Missing
1	1706.00	774.00	2480.00	31.20

Si la recodificamos llamando a la nueva variable P45m seguiremos el proceso siguiente:



Pero al crear la varaible P45m la convierte en variable tipo *factor*. Para convertirla a tipo *double* podemos crear una nueva variable x en blanco con este formato, copiar la información de la columna de la variable P45m, borramos la columna P45m y renombramos la variable x como P45m. A continuación le pedimos la tabla de frecuencia y el descriptivo de la media. Este es el resultado:

Frequencies (P45m)

	Value	# of Cases	%	Cumulative %
1	0	10	0.60	0.60
2	150	28	1.60	2.20
3	450	185	10.80	13.10
4	750	297	17.40	30.50
5	1050	347	20.30	50.80
6	1500	386	22.60	73.40
7	2100	215	12.60	86.00
8	2700	120	7.00	93.10
9	3750	76	4.50	97.50
10	5250	31	1.80	99.40
11	7500	11	0.60	100.00

Case Summary (P45m)

	Valid	Missing	Total	% Missing
1	1706.00	774.00	2480.00	31.20

	Mean
P45m	1500.18

La media de los ingresos de los hogares de la muestra es de 1500 €.

► Ejercicio 15. Propuesto
Recodificar la variable P46 relativa a los ingresos personales con la marca de clase de los intervalos y calcular la media de los ingresos.

2.2.2.2. *Expresiones de transformación*

Veremos a continuación los procedimientos de transformación que implican la realización de un cálculo o una transformación condicional para generar nuevas variables. La utilización de sus comandos implica trabajar con las llamadas expresiones de transformación que especifican la sintaxis de las instrucciones de los comandos de transformación utilizando diferentes tipos de operadores y funciones. En estas expresiones podemos utilizar operadores aritméticos: + – * / ^, constantes, funciones de todo tipo, operadores relacionales: > >= < <= == != y operadores lógicos: & | ! .

2.2.2.3. *Cálculo de variables*

La creación de nuevas variables realizando cálculos es una necesidad constante de todo proceso de análisis de datos cuantitativos. Ya sea para modificar o combinar las variables originales existentes podemos operar infinidad de transformaciones ya sea de naturaleza estadística para acondicionar variables en un análisis, para crear indicadores y nuevas variables cuantitativas, para emplear variables instrumentales, etc.

Los cálculos en R se realizan desde la línea de comandos (o a través de *scripts*). Realizaremos algunos ejercicios de cálculo de variables. En primer lugar podemos plantearnos crear un índice de activismo sociopolítico a partir de las respuestas a la pregunta P14:

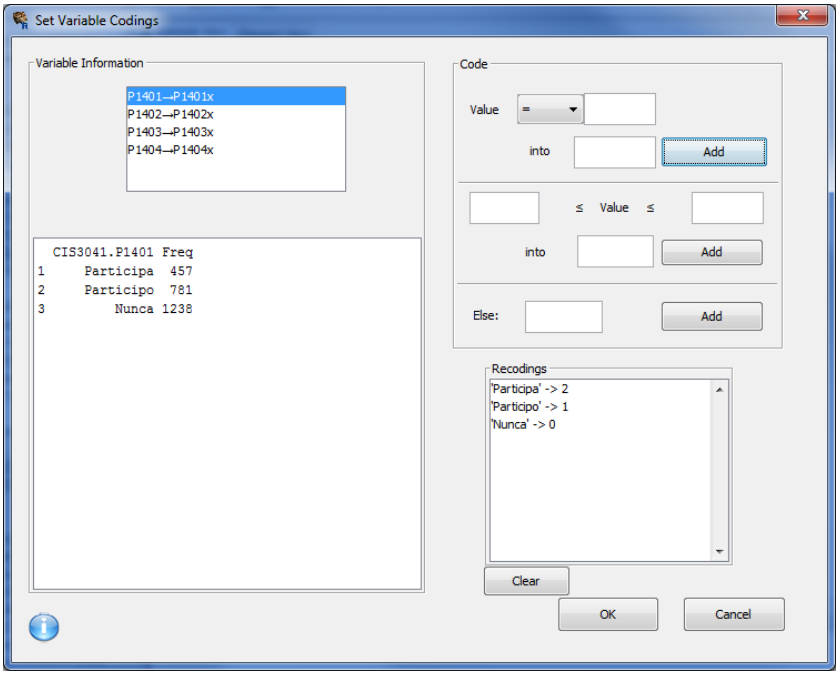
P.14 Existen diversas formas de participación en acciones sociales y políticas que la gente puede llevar a cabo. Por favor, indíqueme para cada una de ellas, si Ud.: ***(MOSTRAR TARJETA D)***.

1. Ha participado durante los últimos doce meses
2. Participó en un pasado más lejano
3. Nunca ha participado

	1	2	3	NC
- Asistir a una manifestación	1	2	3	9 (86)
- Participar en una huelga	1	2	3	9 (87)
- Participar en un foro o en un blog de debate político en Internet	1	2	3	9 (88)
- Firmar una petición /recogida de firmas bien en persona o por Internet	1	2	3	9 (89)

Con los siguientes criterios: puntuar cada forma con 2 si se ha participado últimamente, con 1 si participó en el pasado y con 0 si nunca ha participado. El índice lo construimos con esas puntuaciones en las 4 preguntas sumándolas para cada individuo. El que participe actualmente en todo tendrá un nivel de participación de 8 y el que nunca haya participado en nada de 0. A la nueva variable la llamaremos P14indice.

Teniendo en cuenta los valores actuales de las variables (P1401 a P1404) necesitamos pasar de tipo *factor* a tipo *double* recodificando los valores de las variables como en el caso de la última recodificación comentada en el apartado anterior. Podemos hacerlo para las 4 variables simultáneamente y las llamaremos P1401x a P1404x:



Una vez cambiadas a formato *double*, creamos el índice desde la línea de comandos de la consola de Deducer de la forma siguiente:

```
> CIS3041$P14indice = CIS3041$P1401x + CIS3041$P1402x +
CIS3041$P1403x + CIS3041$P1404x
```

La instrucción contiene a la izquierda el nombre de la nueva variable (P14indice) que se asocia con la matriz de datos CIS3041 (se añadirá como última variable a la matriz de datos) y es el resultado de la expresión de cálculo numérico que implica sumar las 4 variables para cada individuo. Cuando le damos a la tecla <Enter> se crea la variable. Nuestra matriz contendrá una variable más, la última. Hay que tener en cuenta que en la nueva variable algunos individuos son valores perdidos en alguna de las cuatro variables iniciales por lo que no se podrá realizar el cálculo para ellos y serán valores perdidos en la nueva. La tabla de frecuencias de la nueva variable es la siguiente:

Frequencies (P14indice)

	Value	# of Cases	%	Cumulative %
1	0	805	33.00	33.00
2	1	324	13.30	46.20
3	2	417	17.10	63.30
4	3	299	12.20	75.50
5	4	264	10.80	86.30
6	5	127	5.20	91.50
7	6	127	5.20	96.70
8	7	36	1.50	98.20
9	8	44	1.80	100.00

Case Summary (P14indice)

	Valid	Missing	Total	% Missing
1	2443.00	37.00	2480.00	1.50

	Mean
P14indice	2.09

Si calculamos la media se obtiene un valor de 2,09, mucho más cerca de 0 que de 8, indicando un nivel de activismo sociopolítico de la sociedad española en su conjunto relativamente bajo.

► Ejercicio 16. Propuesto

A partir de la pregunta P11 sobre la frecuencia con que se consultan los periódicos, la radio y la televisión para seguir la actualidad política, dando entre 4 y 0 puntos a las frecuencias que van de 1 (Todos los días) a 5 (Nunca) y sumando las puntuaciones para cada individuo.

La operación de tipificación o estandarización de una variable es una transformación que consiste en restar la media a cada puntuación o valor de una variable cuantitativa y dividir por la desviación típica.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Realizamos esta operación con la variable edad (P32). Necesitamos conocer previamente los valores de la media y la desviación ejecutamos el procedimiento Analysis / Descriptives y se obtiene:

	Mean	St. Deviation	Valid N
P32	48.32	17.49	2480

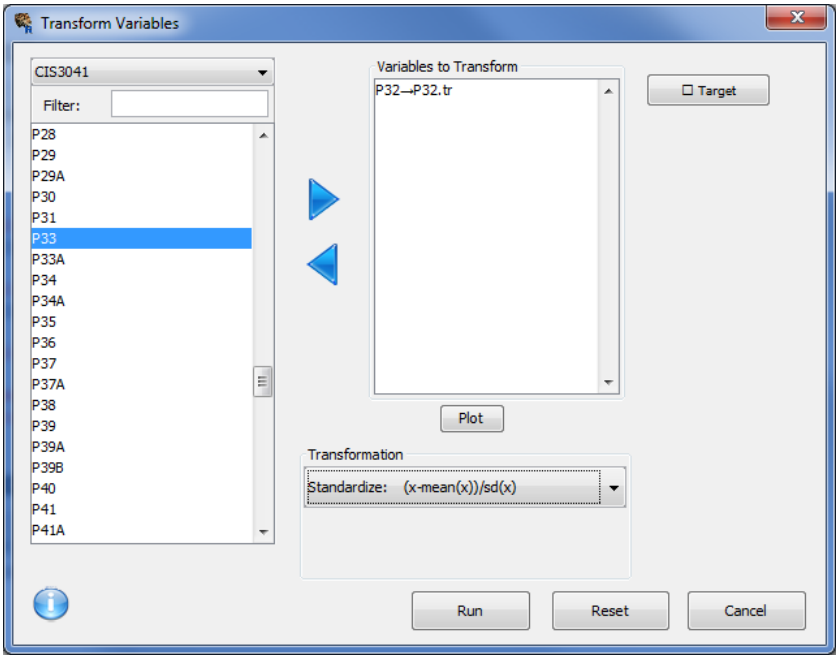
Una vez conocidos los valores de la media y la desviación típica creamos la nueva variable, con el nombre Edadtip, mediante:

```
> CIS3041$Edadtip = (CIS3041$P32-48.32)/17.49
```

Si pedimos los descriptivos de la nueva variable podemos comprobar cómo, salvo decimales, la media es 0 y las desviación típica es 1.

	Mean	St. Deviation	Valid N
Edadtip	-0.000221	1.00	2480

A este mismo resultado se puede llegar a través del menú con Data / Transform, tras elegir la variable P32, pasarla a la derecha y elegir la Transformation Standardize:



Veremos en la matriz de datos añadida al final la variable P32.tr, coindidente con la que creamos anteriormente. A través de estos procedimientos se pueden operar otras transformaciones prestablecidas o incluso proponer la nuestra:
Center: Reescala las variables para que tengan media 0.
Standardize: Reescala las variables para que tengan media 0 y desviación estándar 1.
Robust Standardize: Reescala las variables para que tengan media 0 y desviación absoluta mediana 1.
Range: Transforma la variable para que tome valores entre 0 y 1.
Box-cox: Transforma la variable para intentar obtener una distribución normal.
Rank: Reemplaza los valores por su rango.
Log: Devuelve el logaritmo neperiano (para valores mayores que 0).
Square root: Devuelve la raíz cuadrada.
Absolute value: Devuelve el valor absoluto.
Quantiles: Divide la variable en grupos con el mismo número de observaciones.
Equal width: Divide la variable en grupos con intervalos de la misma amplitud.
Custom: Permite definir transformaciones personalizadas.

Procederemos ahora a la construcción de los indicadores sobre la situación política que elabora el CIS en el Barómetro[32]. Las preguntas de los barómetros de todos los meses

[32] Se puede consultar la metodología para la construcción de indicadores del Barómetro del CIS en la página: http://www.cis.es/cis/opencms/ES/11_barometros/metodologia.html.

relativas a la situación política que se utilizan en la construcción del indicador son la P4 y la P6:

P.4 Y refiriéndonos ahora a la situación política general de España, ¿cómo la calificaría Ud.: muy buena, buena, regular, mala o muy mala?

- Muy buena 1
- Buena 2
- Regular 3
- Mala..... 4 (35)
- Muy mala 5
- N.S. 8
- N.C. 9

P.6 Y, ¿cree Ud. que dentro de un año la situación política del país será mejor, igual o peor que ahora?

- Mejor 1
- Igual 2
- Peor 3 (37)
- N.S. 8
- N.C. 9

El Indicador de la Situación Política Actual (SPA), a partir de la pregunta P4 se define como:

$$SPA = \frac{100.p_1 + 75.p_2 + 50.p_3 + 25.p_4 + 0.p_5}{p_1 + p_2 + p_3 + p_4 + p_5}$$

donde p_1, p_2, p_3, p_4 y p_5 son, respectivamente, los porcentajes de respuesta de las opciones muy buena, buena, regular, mala y muy mala.

El Indicador de Expectativas Políticas (IEP) a partir de la pregunta P6 será:

$$IEP = \frac{100.p_1 + 50.p_2 + 0.p_3}{p_1 + p_2 + p_3}$$

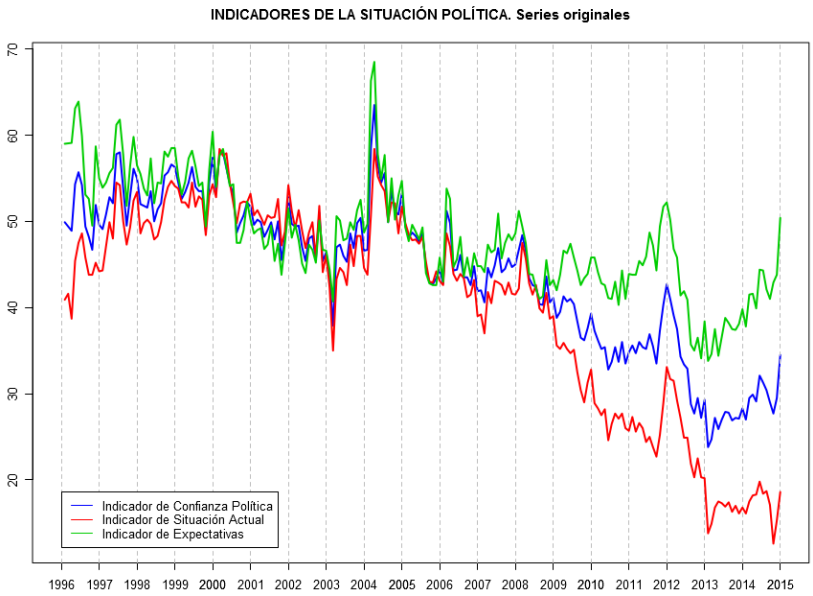
donde p_1, p_2 y p_3 son, respectivamente, los porcentajes de respuesta de las opciones mejor, igual y peor.

Por último, el Indicador de Confianza Política (*ICP*) es la media aritmética de los dos anteriores:

$$ICP = \frac{SPA + IEP}{2}$$

En este caso se trata de indicadores sintéticos que se expresan en un solo valor para el conjunto de la muestra, para después ser comparado a lo largo del tiempo con Barómetros anteriores[33].

[33] Ver http://www.cis.es/cis/export/sites/default/-Archivos/Indicadores/documentos_html/IndiPol.html.



Fuente: CIS

Las frecuencias de ambas variables para octubre de 2014 son:

Frequencies (P4)

	Value	# of Cases	%	Cumulative %
1	Muy buena	2	0.10	0.10
2	Buena	49	2.00	2.10
3	Regular	357	14.90	17.00
4	Mala	769	32.00	49.00
5	Muy mala	1227	51.00	100.00

Frequencies (P6)

	Value	# of Cases	%	Cumulative %
1	Mejor	287	13.30	13.30
2	Igual	1194	55.40	68.70
3	Peor	676	31.30	100.00

Case Summary (P4)

	Valid	Missing	Total	% Missing
1	2404.00	76.00	2480.00	3.10

Case Summary (P6)

	Valid	Missing	Total	% Missing
1	2157.00	323.00	2480.00	13.00

Para obtener los 3 indicadores utilizaremos la línea de comandos de la consola como "calculadora":

```
> SPA=((100*0.1)+(75*2.0)+(50*14.9)+(25*32.0)+(0*51.0))/100
> SPA
[1] 17.05
> IEP=((100*13.3)+(50*55.4)+(0*31.3))/100
> IEP
[1] 41
> ICP=(SPA+IEP)/2
> ICP
[1] 29.025
```

2.2.2.4. *Transformaciones condicionales*

Para finalizar este recorrido por la transformación de las variables trabajaremos con un procedimiento de primera necesidad en el trabajo de análisis de la información cuantitativa: la creación de variables con transformaciones condicionales. Son

situaciones donde se establecen determinadas condiciones en las características de las unidades y en función de su cumplimiento según una expresión lógica (verdadero o falso / perdido) asigna un valor a través de una expresión (dando el valor en concreto o ejecutando una fórmula de cálculo). La transformación condicional se puede utilizar en diversos comandos, pero nos detendremos sobre todo en el comando **`ifelse`**.

El comando **`ifelse`** que tiene la forma general siguiente: ifelse(test, yes, no). Se evalúa un condición (test) y si es verdadera se ejecuta una transformación (yes), en caso contrario se ejecuta otra trasformación o acción (no).

A través de las transformaciones condicionales se construyen las variables tipológicas que combinan simultáneamente características de diversas variables (espacio de atributos) para definir diversos tipos. Es el caso de la construcción de la variable de clase social, del estilo de vida, de tipo de consumidor, etc.

Para ilustrar la utilización de ese procedimiento con R crearemos una variable (tipológica) de movilidad ocupacional intergeneracional a partir de relacionar el nivel ocupacional del padre con el alcanzado por el hijo/a. Las variables ocupacionales son respectivamente OCUPAPAD y OCUMAR11. Como paso previo pediremos la tabla de contingencia que cruza ambas variables (Analysis / Contingency Tables) para visualizar la información que se trabaja, ilustrar el procedimiento y luego poder verificar la creación de la nueva variable. Por convención, en filas se coloca el origen social del padre y en columnas el del hijo/a. La tabla es la siguiente:

		OCUMAR11 Ocupación del hijo/a									
		1	2	3	4	5	6	7	8	9	Total
OCUPAPAD	1	13	19	13	3	12	0	2	1	6	69
Ocupación	2	4	75	19	4	18	0	10	5	2	137
del padre	3	10	34	58	13	46	1	8	15	10	195
	4	1	7	9	9	14	1	3	1	4	49
	5	18	34	36	15	98	6	26	11	28	272
	6	7	26	35	9	80	84	73	60	50	424
	7	12	44	64	15	121	9	121	48	70	504
	8	7	33	48	11	79	7	50	91	29	355
	9	2	12	8	5	25	7	24	20	53	156
Total		74	284	290	84	493	115	317	252	252	2161

1 Directores y gerentes; 2 Técnicos y profesionales científicos e intelectuales; 3 Técnicos; profesionales de apoyo; 4 Empleados contables, administrativos y otros empleados de oficina; 5 Trabajadores de los servicios de restauración, personales, protección y vendedores; 6 Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero; 7 Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción, excepto operadores de instalación; 8 Operadores de instalaciones y maquinaria, y montadores; 9 Ocupaciones elementales

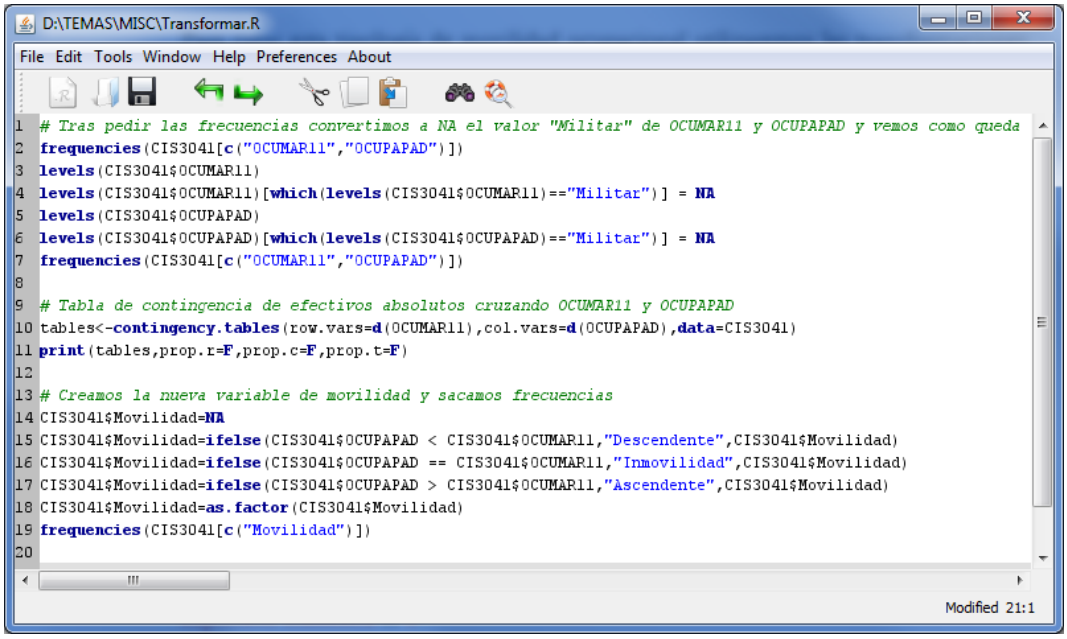
La diagonal (en azul) define la inmovilidad o la reproducción social ocupacional donde el origen ocupacional del padre es el mismo que el del hijo/a. Los valores del triángulo inferior (en verde) corresponden a la movilidad ascendente, los hijos/as tienen un nivel ocupacional más alto que los padres. Finalmente el triángulo superior (en rojo) corresponden a la movilidad descendente, los hijos/as tiene menor nivel ocupacional.

Para crear esta tipología de movilidad ocupacional utilizaremos las transformaciones condicionales. En este caso establecemos 3 condiciones[34]:

- Si OCUPAPAD < OCUMAR11 entonces se da movilidad descendente
- Si OCUPAPAD = OCUMAR11 entonces se da inmovilidad
- Si OCUPAPAD > OCUMAR11 entonces se da movilidad ascendente

Todos los casos que no cumplan estas condiciones, es decir, los casos que corresponden valores perdidos de ambas variables, pasarán a ser valores perdidos del sistema. Para obtener la tabla anterior de 9 por 9 categorías debemos considerar al valor 10 "Militar" como valor perdido.

Para obtener la tipología de movilidad ocupacional con R ejecutaremos instrucciones en el lenguaje de comandos a través de la elaboración de un programa de sintaxis (*script*). Para crear el archivo de sintaxis abrimos con File / New Document y escribiremos las siguientes instrucciones que pasamos a comentar[35]:



Se solicitan en primer lugar las tablas de frecuencias de las dos variables con el comando **`frequencies`**,[36] que solamente funciona con Deducer abierto o si se tiene cargada la librería, pues no es un comando de la librería base de R. El comando **`levels`** permite ver los atributos de una variable y también cambiarlos, como en este caso donde el valor Militar se convierte en NA en la dos variables. Se utiliza además el comando **`which`** que buscar el valor que corresponde al atributo Militar en la variable. Para ejecutar las instrucciones de la sintaxis creada se selecciona y se teclea <CTRL>+<R>. Las frecuencias de las variables son:

[34] Como los valores van de 1, mayor nivel ocupacional, a 9, menor nivel, el sentido de la comparación es el inverso: un valor mayor entre origen y destino es movilidad descendente y un valor menor ascendente.

[35] Las instrucciones se encuentran en el archivo Transformar.R.

[36] Las variables aparecen asociadas al data frame al cual pertenecen CIS3014 para indicar en qué fichero está la variable y dónde se ha de guardar si se crea una nueva. En R existen dos comandos: **`attach`** y **`detach`** que permiten gestionar este aspecto, el primero evitar escribir constantemente el nombre de la matriz estableciedo la base de datos por defecto y el segundo anula la acción.

Frequencies (OCUMAR11)

	Value	# of Cases	%	Cumulative %
1	Director	84	3.40	3.40
2	Tecnico	309	12.70	16.10
3	Apoyo	325	13.30	29.50
4	Administrativo	100	4.10	33.60
5	Servicios	559	22.90	56.50
6	Cualificado agricola	132	5.40	61.90
7	Cualificado industria	359	14.70	76.70
8	Operadores	274	11.20	87.90
9	Elemental	294	12.10	100.00

Frequencies (OCUPAPAD)

	Value	# of Cases	%	Cumulative %
1	Director	72	3.30	3.30
2	Tecnico	138	6.30	9.60
3	Apoyo	199	9.10	18.70
4	Administrativo	50	2.30	20.90
5	Servicios	279	12.70	33.70
6	Cualificado agricola	429	19.60	53.20
7	Cualificado industria	512	23.30	76.60
8	Operadores	358	16.30	92.90
9	Elemental	156	7.10	100.00

Case Summary (OCUMAR11)

	Valid	Missing	Total	% Missing
1	2436.00	44.00	2480.00	1.80

Case Summary (OCUPAPAD)

	Valid	Missing	Total	% Missing
1	2193.00	287.00	2480.00	11.60

A continuación se pide la tabla de contingencia, este también es un comando propio de la librería Deducer. Su ejecución genera este resultado:

OCUMAR11 by OCUPAPAD across levels of

		OCUPAPAD									
OCUMAR11		Director	Tecnico	Apoyo	Administrativo	Servicios	Cualificado agricola	Cualificado industria	Operadores	Elemental	Row Total
Director	Count	13	4	10	1	18	7	12	7	2	74
Tecnico	Count	19	75	34	7	34	26	44	33	12	284
Apoyo	Count	13	19	58	9	36	35	64	48	8	290
Administrativo	Count	3	4	13	9	15	9	15	11	5	84
Servicios	Count	12	18	46	14	98	80	121	79	25	493
Cualificado agricola	Count	0	0	1	1	6	84	9	7	7	115
Cualificado industria	Count	2	10	8	3	26	73	121	50	24	317
Operadores	Count	1	5	15	1	11	60	48	91	20	252
Elemental	Count	6	2	10	4	28	50	70	29	53	252
Column Total		69	137	195	49	272	424	504	355	156	**2161**

Finalmente se procede a la construcción de la nueva variable que llamaremos Movilidad. Empezamos creando la variable con todos los valores perdidos y los modificamos a continuación según las condiciones que comentamos más arriba que definen los tres tipos de movilidad. La primera de ellas establece con el comando **`ifelse`** la condición que se ha de satisfacer para asignar el valor Descendente a un individuo en la nueva variable (movilidad descendente), OCUPAPAD < OCUMAR11. Si se da la condición se asigna el valor Descendente a todos los casos que la cumplan, en caso contrario el valor que tenga en la variable inicialmente, es decir, NA. Las otras dos condiciones de forma equivalente establecen la inmovilidad, OCUPAPAD == OCUMAR11 y la movilidad ascendente, OCUPAPAD > OCUMAR11. Para acabar se cambia el tipo de variable creada, se convierte del formato *character* con el que se genera a *factor*, y cambiamos el orden de las etiquetas para convertirla en una variable *factor ordered*. La tabla de frecuencias que se obtiene es la siguiente:

Frequencies (Movilidad)

	Value	# of Cases	%	Cumulative %
1	Descendente	631	29.20	29.20
2	Inmovilidad	602	27.90	57.10
3	Ascendente	928	42.90	100.00

Case Summary (Movilidad)

	Valid	Missing	Total	% Missing
1	2161.00	319.00	2480.00	12.90

Como se puede observar destaca la movilidad ocupacional absoluta ascendente (43%) como resultado del proceso de cambios que ha experimentado la sociedad española desde el periodo de industrialización a la fase postindustrial actual.

► **Ejercicio 17. Propuesto**
Realizar un análisis de la relación entre nivel educativo (variable ESTUDIOS) y la ocupación (variable OCUMAR11) de las personas entrevistadas. Proponer la creación de una tipología empírica que las relacione a partir de las frecuencias observadas en la tabla de contingencia.

► **Ejercicio 18. Propuesto**
Crear una variable tipológica que relacione el dinero y la felicidad, considerando las variables Escala de felicidad personal (P30) e Ingresos personales (P46). Para ello recodificar previamente cada una de las variables en tres categorías: feliz, ni feliz ni feliz, infeliz para la felicidad, y rico, ni rico ni pobre y pobre para los ingresos. Responder a la pregunta: ¿hasta qué punto el dinero hace la felicidad?

Como hemos ido viendo a lo largo de este apartado, la realización de transformaciones con las variables implica modificar o crear otras nuevas que van ampliando nuestro fichero datos como pusimos de manifiesto al inicio de este capítulo al hablar del proceso de datos. Ello implica gestionar cómo guardar estos datos. Una práctica recomendable es mantener una copia de la fuente de datos original y crear la matriz ampliada guardándola con otro nombre. En nuestro caso todas las variables que hemos ido generando se encuentran en la matriz CIS3041+.rda.

Conviene observar también que los datos generados se han obtenido en general desde el menú en una dinámica de trabajo interactiva lo que puede representar una limitación de cara a replicar el trabajo realizado. Para volver a realizar los ejercicios vistos disponemos del propio manual, pero en la práctica de la investigación, revisar o rehacer la generación de los datos y su análisis requiere registrarlo. Una forma de hacerlo es guardar sistemáticamente los archivos de resultados que contienen la sintaxis y los resultados de su ejecución. Pero volver a ejecutarlos por el menú para traducir aquellos comandos y resultados puede resultar complicado, largo y laborioso. La alternativa es guardar archivos de sintaxis con todas las tareas realizadas que al ser ejecutados de nuevo, en cuestión de segundos, generan todo el trabajo de horas que representó cuando se diseñaron originalmente. Así hemos trabajado nosotros y hemos guardado todas las transformaciones que se han visto en el capítulo en el programa de sintaxis Transformar.R que se puede consultar en la página web de este capítulo.

3. Bibliografía

Badiella, Ll. et al. (2015). *Manual de Introducción a Deducer: una interfaz gráfica para usuarios de R*. Bellaterra (Cerdanyola del Vallès). Servei d'Estadística Aplicada de la Universitat Autònoma de Barcelona. 5ª edición. http://sct.uab.cat/estadistica/sites/sct.uab.cat.estadistica/files/Manual%20curs%20Deducer.pdf

Bouso, J. (2013). *El paquete estadístico R*. Madrid: Centro de Investigaciones Sociológicas.

Chapman, G. (2012). *Deducer Quick Start Guide*. Exploring Computer Science. National Science Foundation. http://www.exploringcs.org/wp-content/uploads/2010/08/Deducer-Quick-Start-Guide.pdf

Domínguez, M.; Simó, M. (2003). *Tècniques d'Investigació Social Quantitatives*. Barcelona: Edicions Universitat de Barcelona. Metodologia, 13.

Dalgaard, P. (2008). *Introductory Statistics with R*. New York: Springer.

Díaz de Rada, V. (2002). *Técnicas de análisis de datos para investigadores sociales. Aplicaciones prácticas con SPSS para Windows*. Madrid: RA-MA.

Díaz de Rada, V. (2009). *Análisis de datos de encuesta*. Barcelona: Editorial UOC.

Fachelli, S.; López-Roldán, P. (2013). ¿Somos más móviles? Incluyendo a la mitad invisible. *XI Congreso Español de Sociología*, Madrid 10-12 de julio de 2013. http://www.fes-web.org/uploads/files/modules/congress/11/papers/1923.pdf.

Fachelli, S.; López-Roldán, P. (2015). ¿Somos más móviles incluyendo a la mitad invisible? Análisis de la movilidad social intergeneracional en España en 2011. *Revista Española de Investigaciones Sociológicas*, 150.

IBM Corporation (2013). *IBM SPSS Statistics 22 Command Syntax Reference*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/en/client/Manuals/IBM_SPSS_Statistics_Command_Syntax_Reference.pdf.

IBM Corporation (2015a). *IBM SPSS Statistics 22 Core System. Guía del usuario*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_User_Guide.pdf.

IBM Corporation (2015b). *IBM SPSS Statistics Base 22*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf.

IBM Corporation (2015c). *Guía breve de IBM SPSS Statistics 22*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf.

Lizasoaín, L.; Joaristi, L. (2003). *Gestión y análisis de datos con SPSS: versión 11*. Madrid: Paraninfo.

López-Roldán, P. (2014). *Análisis de datos con SPSS*. En P. López-Roldán, *Recursos per a la investigación social*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. http://ddd.uab.cat/record/89349

Murillo Torrecilla, F. J.; Martínez-Garrido, C. (2012). *Análisis de datos cuantitativos con SPSS en investigación socioeducativa*. Madrid: Servicio de Publicaciones de la Universidad Autónoma de Madrid.

Muenchen, R. A. (2011). *R fos SAS and SPSS Users*. New York: Springer. 2ª edición.

Pardo, A.; Ruiz, M. A. (2005). *Análisis de datos con SPSS 13*. Madrid: McGraw-Hill.

Pardo, A.; Ruiz, M. A. (2009). *Gestión de datos con SPSS Statistics*. Madrid: Síntesis.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing Vienna, Austria. ISBN: 3-900051-07-0. http://www.r-project.org/.

Rial, A.; Varela, J.; Rojas, A. J. (2001). *Depuración y análisis preliminares de datos en SPSS*. Madrid: RA-MA.

Spector, Ph. (2008). *Data Manipulation with R*. New York: Springer.