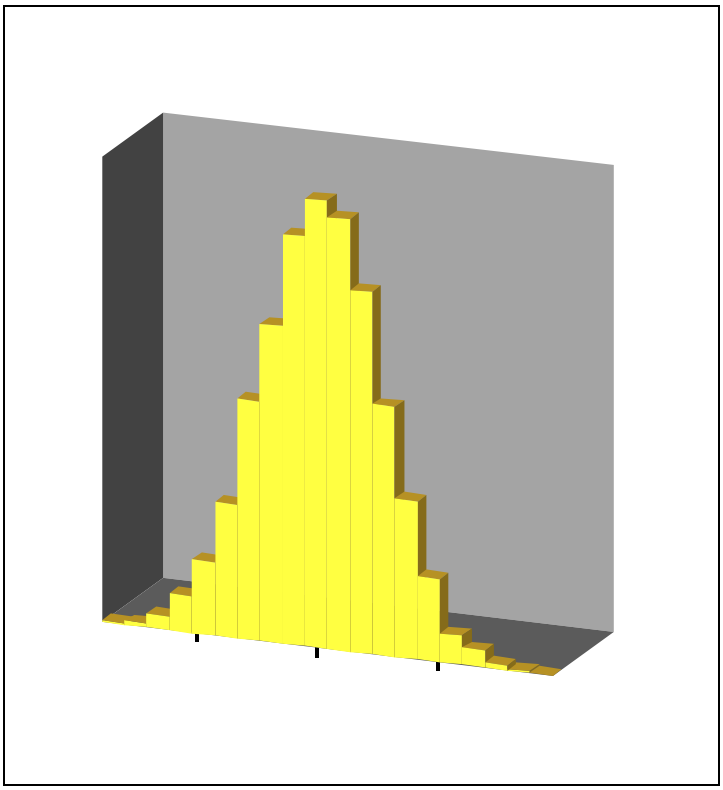


TRATAMIENTO DE DATOS EN LA INVESTIGACIÓN PSICOLÓGICA CON SPSS

PROGRAMA DE DOCTORADO
DEL DEPARTAMENTO DE METODOLOGÍA
DE LAS CIENCIAS DEL COMPORTAMIENTO

Enrique Moreno González
Departamento de Metodología
Facultad de Psicología
UNED

ÍNDICE

Presentación general del programa SPSS

Introducción

El presente curso tiene como objetivo acercar al usuario al manejo del *software* de análisis estadístico SPSS, acrónimo de *Statistical Package for Sciences Socials* (Paquete Estadístico para las Ciencias Sociales), en sus aspectos más básicos, los que se refieren al tratamiento general de datos y los relativos a ciertos análisis estadísticos considerados simples, es decir, descripción general de cualquier tipo de variable estadística y evaluación de relaciones entre dos variables, dejando para un futuro análisis más complejos, de carácter multivariante, que también pueden realizarse con este programa.

En primer lugar, y antes de comenzar a desarrollar los contenidos específicos de este curso, daremos un breve paseo por las versiones anteriores de SPSS para ver la evolución que ha experimentado hasta llegar a la actual versión, la 10.0.

Para el desarrollo del curso se emplean los mismos archivos que SPSS incluye en el CD-ROM en el que se distribuye el programa. En cada momento haremos mención al archivo con el que vamos a trabajar. Todos los archivos, una vez instalado SPSS en el ordenador, se encuentran en la misma ruta C:\Archivos de Programa\SPSS\

Antes de comenzar, expreso el deseo de que este manual os sirva de guía para moveros con sencillez por las pantallas del programa y realizar los procedimientos de análisis más básicos. Por supuesto, aceptaré todos los comentarios que tengáis a bien hacerme para mejorar este manual en la medida de lo posible.

Breve historia del SPSS

A finales de la decada de los 80 SPSS desarrolló un programa de análisis estadístico para su ejecución en los ordenadores personales, bajo el entorno operativo MS-DOS. Hasta entonces había versiones del mismo para grandes plataformas (*mainframe*), que habitualmente conformaban los equipos de los centros de cálculo de las universidades y laboratorios de investigación. Para llevar a cabo los análisis era preciso escribir las instrucciones en un lenguaje específico de SPSS, con una sintaxis particular. Este lenguaje que soportaba SPSS para grandes equipos se ha transmitido, con ligeras variaciones, a las sucesivas versiones para ordenadores personales, tanto en el entorno MS-DOS como en el de WINDOWS, aunque en este último pueda llegar a pasar desapercibido para el principiante.

Como muestra veamos cómo se podría obtener una distribución de frecuencias de una variable V1 contenida en un archivo de datos con tres variables (V1, V2 y V3). Las instrucciones serían las siguientes:

```
DATA LIST FILE ='C:\CURSPSS\ARCHIV1.DAT'/ V1 1-3 V2 5-6 V3 8-20(a).
FRECUENCIES V1/ STATISTICS = NONE.
```

En términos llanos, estas dos sentencias podrían traducirse así:

> "... *leer el archivo de datos en formato ASCII, ARCHIV1.DAT (DATA LIST FILE) ubicado en el directorio CURSPSS de la unidad C, el cual contiene tres variable: V1 con tres dígitos que ocupa las columnas 1 a 3; V2 que ocupan las columnas 5 y 6; y V3 que ocupan las columnas 8 a 20 es una variable de cadena, tal como se especifica por la letra a dentro del paréntesis*".

Posteriormente, confeccionar una distribución de frecuencias de la variable V1, y no calcular estadísticos (STATISTICS = NONE)..."

De esta forma, escribiendo los procedimientos adecuados, se obtenían todos los análisis que incorporaba el SPSS.

Como se ha dicho, esta sintaxis se mantiene, ampliada, en todas las sucesivas versiones que han salido al mercado, para ser implementadas en los ordenadores personales. No obstante, ya en la versión 4 para DOS, aparecieron los primeros menús de ayuda en línea mediante los cuales se podían obtener los mismos resultados sin tener que escribir los procedimientos. De esta forma se elegían en dichos menús los procedimientos que se iban a utilizar y el programa escribía en un editor de texto (REVIEW) la sentencia adecuada en función del procedimiento elegido; SPSS empezaba a dulcificar el *interface* de usuario.

En estas versiones de SPSS para DOS había un déficit importante, que era el asunto de los gráficos. Para obtenerlos era preciso tener grabado en el ordenador algún software de gráficos, y configurar SPSS para que pudiera trabajar con ese software en cuestión (por defecto solía trabajar con HARVARD-GRAPHICS), lo cual, para un usuario poco avezado, podía suponer un problema añadido.

Este inconveniente ha sido subsanado en las versiones para Windows, y SPSS ya dispone de un software de generación de gráficos integrado en la aplicación y con las opciones propias de los editores de gráficos.

Después de este breve repaso por la historia del SPSS vamos a comenzar el curso de la manera más directa posible: realizando una sesión completa de trabajo, que nos permitirá obtener una visión global de las características más notables de la aplicación. Posteriormente, en los siguientes capítulos, profundizaremos en cada una de las operaciones básicas y procedimientos que se pueden realizar, desde la edición de datos a la elaboración de análisis estadísticos, pasando por el tratamiento de esos datos (creación de nuevas variables, transformación de variables, ordenación, ponderación, selección, etc.).

Comencemos pues.

1. Primera sesión con SPSS

1.1 Aspectos básicos

Cualquier sesión tipo se puede resumir en cuatro grandes apartados:

- Lectura de un conjunto de datos
- Selección del Procedimiento
- Selección de Variables
- Examen de Resultado

Pero antes... antes hemos de entrar en SPSS para poder llevar a cabo esta primera sesión. Para ello hay dos maneras de proceder: 1) Desde el menú Programas que se despliega a pulsar el botón **Inicio** se accede al programa SPSS, de la misma manera que se accede a cualquier programa que opere bajo el sistema operativo de Windows, bien en la versión 95 en la 98 o en la 2000; 2) A través de un Icono de Acceso Directo que hayamos creado previamente en el Escritorio o en la barra de accesos rápidos situada en la parte inferior de la pantalla, por el procedimiento habitual de creación de estos tipos de accesos directos1. En ambos casos el resultado es el mismo: se accede al programa, directamente al **Editor de Datos**, cuya apariencia es la que se muestra en la Figura 1.1.

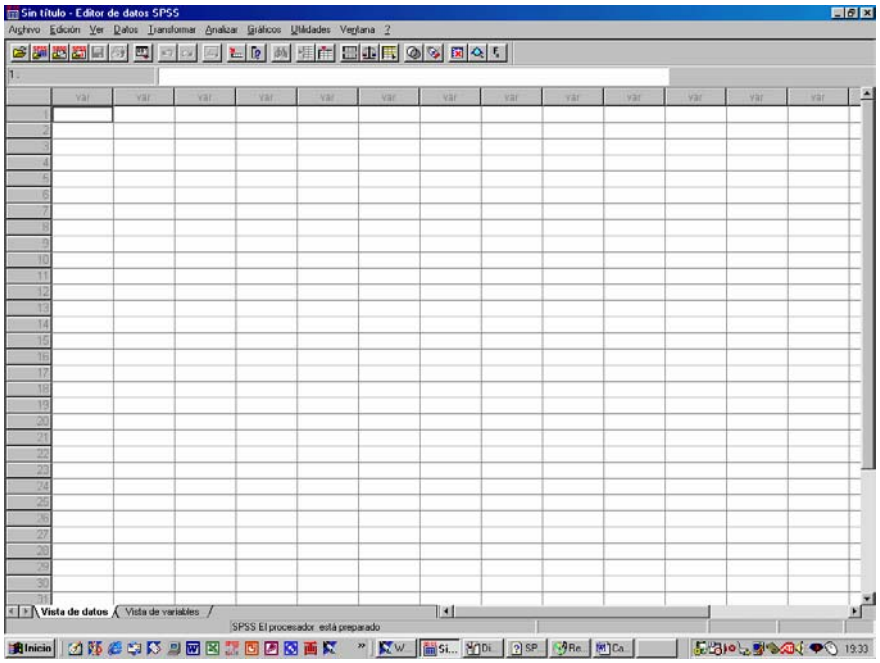


Fig. 1.1 Editor de datos de SPSS, sin datos

1 La diferencia entre un icono de acceso directo en el escritorio y otro en la barra de acceso rápido está en que en el escritorio, si no se ha modificado las opciones de carpeta del panel de control, hay que hacer doble clic para acceder al programa y en la barra sólo un clic.

Es en esta pantalla en la que se va a desarrollar buena parte de las sesiones con SPSS. Aquí es donde grabaremos los datos registrados en el desarrollo de nuestros trabajos, o donde se mostrarán los datos ya grabados en archivos cuando queramos someterlos a los procedimientos de análisis de SPSS.

El aspecto del editor de datos es el propio de una rejilla de filas y columnas cuya intersección conforman las celdillas de la misma -cada celdilla un dato-, similar a la que dispone cualquier hoja de cálculo. En este primera sesión vamos a utilizar los datos previamente almacenados en un archivo, por lo que el primer paso es leer esos datos. Para ello se puede emplear dos maneras alternativas: la primera es a través de la opción **Archivo** del menú principal, sub-opción **Abrir**. La otra alternativa, más inmediata, es pulsar, en los iconos que aparecen debajo del menú general, el correspondiente a **Abrir archivo**  . En ambos casos, se accede a una ventana como la de la Figura 1.2.

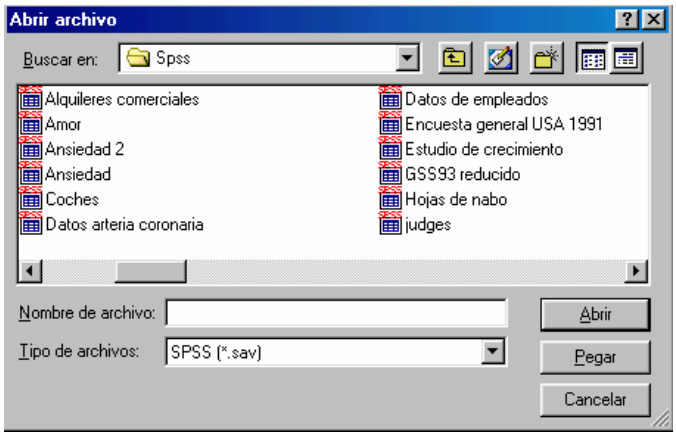
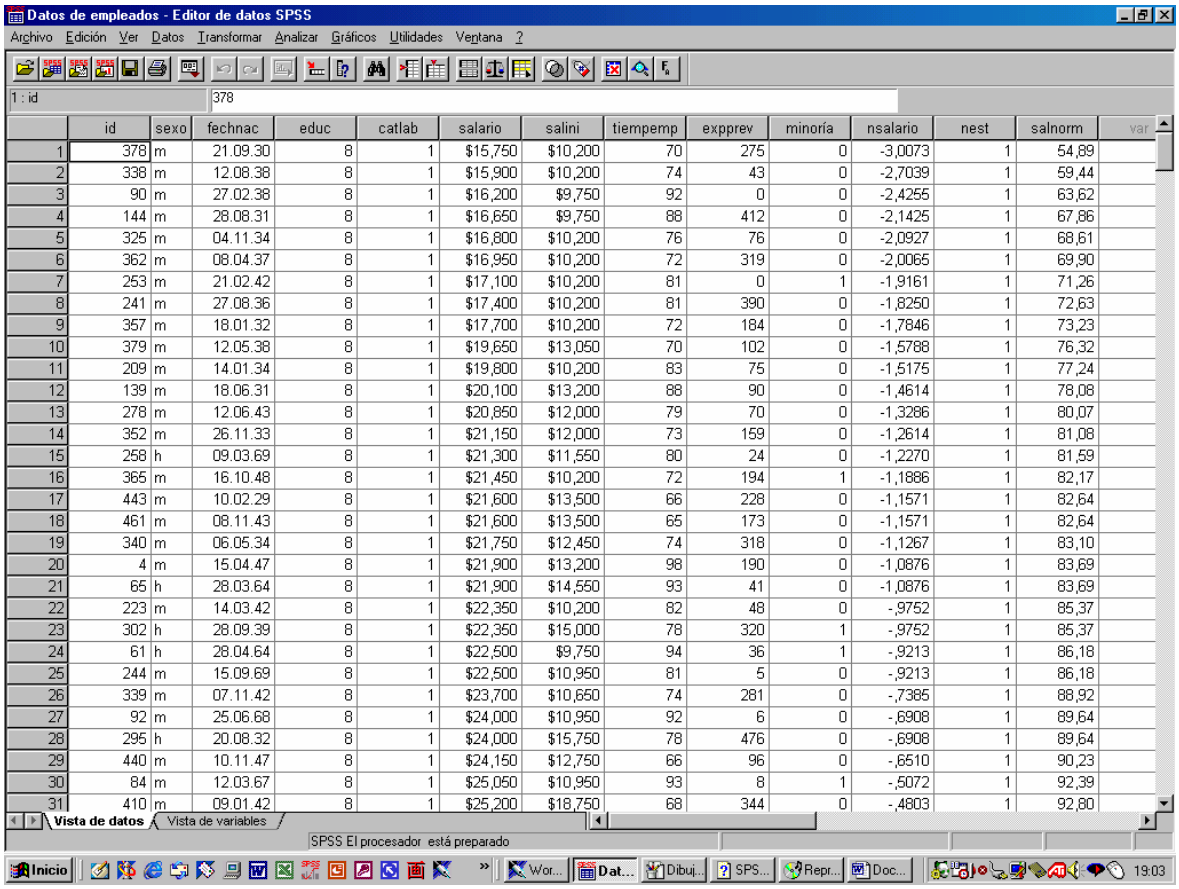


Figura 1.2 Cuadro de diálogo de *Abrir archivo*

Por defecto, sólo se lista los archivos de datos generados y guardados previamente por SPSS, que en las versiones para Windows tienen la extensión SAV, aunque SPSS puede leer datos grabados en diferentes formatos (ASCII, dBASE, Excel, etc.), y por supuesto los archivos generados por las anteriores versiones del programa, que se identifican por la extensión SYS.

Para abrir un archivo de datos, basta hacer doble clic con el botón izquierdo del ratón en el mismo y se incorpora al Editor de datos. El aspecto del editor una vez leído el archivo (en este caso el archivo *Datos de empleados*) es el que se ve en la Figura 1.3.

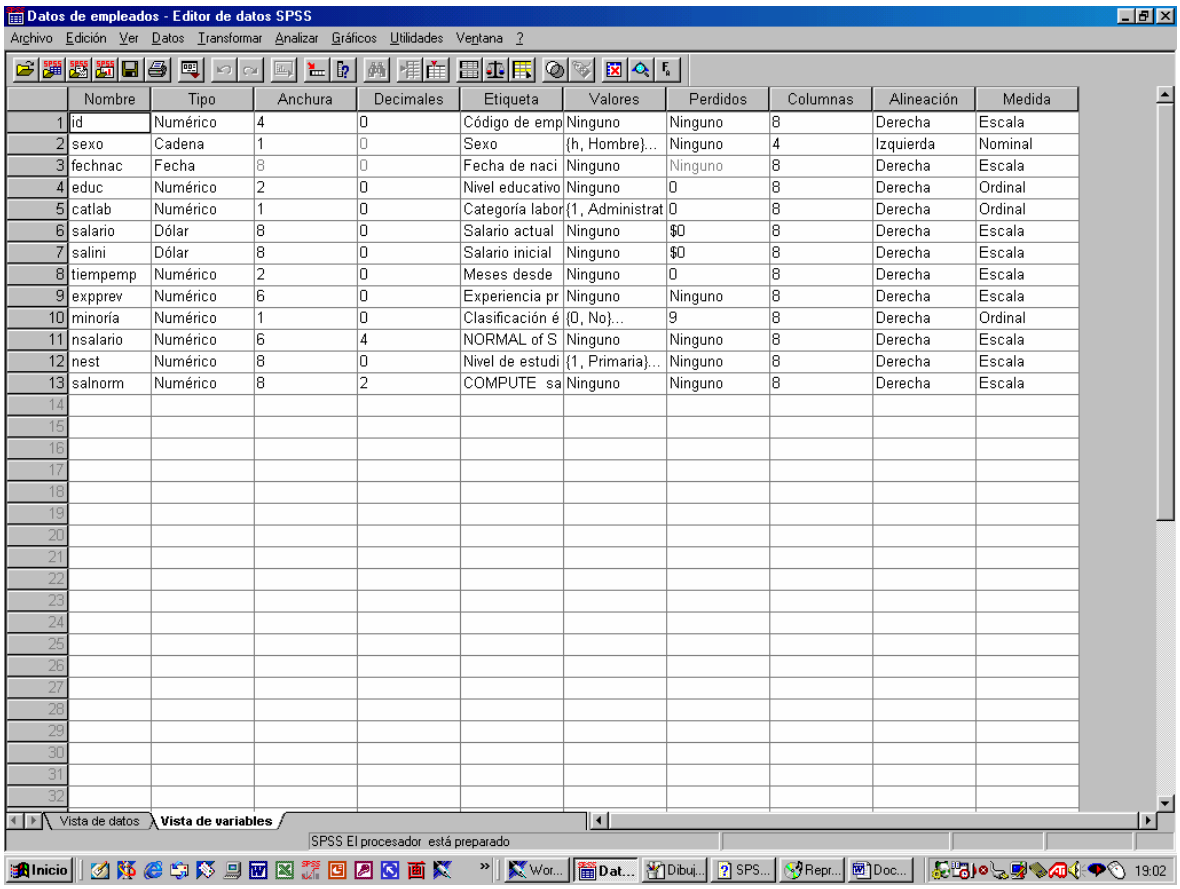


	id	sexo	fechnac	educ	catlab	salario	salini	tiempemp	expprev	minoría	nsalario	nest	salnorm	var
1	378	m	21.09.30	8	1	$15,750	$10,200	70	275	0	-3,0073	1	54,89	
2	338	m	12.08.38	8	1	$15,900	$10,200	74	43	0	-2,7039	1	59,44	
3	90	m	27.02.38	8	1	$16,200	$9,750	92	0	0	-2,4255	1	63,62	
4	144	m	28.08.31	8	1	$16,650	$9,750	88	412	0	-2,1425	1	67,86	
5	325	m	04.11.34	8	1	$16,800	$10,200	76	76	0	-2,0927	1	68,61	
6	362	m	08.04.37	8	1	$16,950	$10,200	72	319	0	-2,0065	1	69,90	
7	253	m	21.02.42	8	1	$17,100	$10,200	81	0	1	-1,9161	1	71,26	
8	241	m	27.08.36	8	1	$17,400	$10,200	81	390	0	-1,8250	1	72,63	
9	357	m	18.01.32	8	1	$17,700	$10,200	72	184	0	-1,7846	1	73,23	
10	379	m	12.05.38	8	1	$19,650	$13,050	70	102	0	-1,5788	1	76,32	
11	209	m	14.01.34	8	1	$19,800	$10,200	83	75	0	-1,5175	1	77,24	
12	139	m	18.06.31	8	1	$20,100	$13,200	88	90	0	-1,4614	1	78,08	
13	278	m	12.06.43	8	1	$20,850	$12,000	79	70	0	-1,3286	1	80,07	
14	352	m	26.11.33	8	1	$21,150	$12,000	73	159	0	-1,2614	1	81,08	
15	258	h	09.03.69	8	1	$21,300	$11,550	80	24	0	-1,2270	1	81,59	
16	365	m	16.10.48	8	1	$21,450	$10,200	72	194	1	-1,1886	1	82,17	
17	443	m	10.02.29	8	1	$21,600	$13,500	66	228	0	-1,1571	1	82,64	
18	461	m	08.11.43	8	1	$21,600	$13,500	65	173	0	-1,1571	1	82,64	
19	340	m	06.05.34	8	1	$21,750	$12,450	74	318	0	-1,1267	1	83,10	
20	4	m	15.04.47	8	1	$21,900	$13,200	98	190	0	-1,0876	1	83,69	
21	65	h	28.03.64	8	1	$21,900	$14,550	93	41	0	-1,0876	1	83,69	
22	223	m	14.03.42	8	1	$22,350	$10,200	82	48	0	-,9752	1	85,37	
23	302	h	28.09.39	8	1	$22,350	$15,000	78	320	1	-,9752	1	85,37	
24	61	h	28.04.64	8	1	$22,500	$9,750	94	36	1	-,9213	1	86,18	
25	244	m	15.09.69	8	1	$22,500	$10,950	81	5	0	-,9213	1	86,18	
26	339	m	07.11.42	8	1	$23,700	$10,650	74	281	0	-,7385	1	88,92	
27	92	m	25.06.68	8	1	$24,000	$10,950	92	6	0	-,6908	1	89,64	
28	295	h	20.08.32	8	1	$24,000	$15,750	78	476	0	-,6908	1	89,64	
29	440	m	10.11.47	8	1	$24,150	$12,750	66	96	0	-,6510	1	90,23	
30	84	m	12.03.67	8	1	$25,050	$10,950	93	8	1	-,5072	1	92,39	
31	410	m	09.01.42	8	1	$25,200	$18,750	68	344	0	-,4803	1	92,80	

Figura 1.3 Ventana de datos en el Editor de Datos con el archivo *Datos de empleados*

Las variables estadísticas grabadas en el archivo, se trasladan al editor de datos con la misma disposición: cada variable en una columna y cada caso u observación en una fila.

El Editor de Datos tiene dos pantallas. En la primera, etiquetada en la pestaña inferior izquierda como **Vista de datos**, están los datos tal como se muestra en la Figura 1.3; en la otra, etiquetada como **Vista de variables**, se definen las variables: nombre, tipo, etc. Esta ventana es similar a la de definición de campos del programa Microsoft Acces, y su aspecto es el que se muestra en la Figura 1.4.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	id	Numérico	4	0	Código de emp	Ninguno	Ninguno	8	Derecha	Escala
2	sexo	Cadena	1	0	Sexo	{h, Hombre}...	Ninguno	4	Izquierda	Nominal
3	fechnac	Fecha	8	0	Fecha de naci	Ninguno	Ninguno	8	Derecha	Escala
4	educ	Numérico	2	0	Nivel educativo	Ninguno	0	8	Derecha	Ordinal
5	catlab	Numérico	1	0	Categoría labor	{1, Administrat	0	8	Derecha	Ordinal
6	salario	Dólar	8	0	Salario actual	Ninguno	$0	8	Derecha	Escala
7	salini	Dólar	8	0	Salario inicial	Ninguno	$0	8	Derecha	Escala
8	tiempemp	Numérico	2	0	Meses desde	Ninguno	0	8	Derecha	Escala
9	expprev	Numérico	6	0	Experiencia pr	Ninguno	Ninguno	8	Derecha	Escala
10	minoría	Numérico	1	0	Clasificación é	{0, No}...	9	8	Derecha	Ordinal
11	nsalario	Numérico	6	4	NORMAL of S	Ninguno	Ninguno	8	Derecha	Escala
12	nest	Numérico	8	0	Nivel de estudi	{1, Primaria}...	Ninguno	8	Derecha	Escala
13	salnorm	Numérico	8	2	COMPUTE sa	Ninguno	Ninguno	8	Derecha	Escala

Figura 1.4. Ventana de definición de variables en el Editor de Datos

1.2 Definición de Variables

La definición de variables se efectúa en la ventana correspondiente a la Vista de variables en el Editor de datos. A continuación se dan una serie de directrices.

Para los nombres de variable se aplican las siguientes normas:

- El nombre debe comenzar por una letra. Los demás caracteres pueden ser letras, dígitos, puntos o los símbolos @, #, _ o $.
- Los nombres de variable no pueden terminar en punto.
- Se deben evitar los nombres de variable que terminan con subrayado (para evitar conflictos con las variables creadas automáticamente por algunos procedimientos).
- La longitud del nombre no debe exceder los ocho caracteres.
- No se pueden utilizar espacios en blanco ni caracteres especiales (por ejemplo, !, ?, ’ y *).
- Cada nombre de variable debe ser único; no se permiten duplicados. Los nombres de variable no distinguen mayúsculas de minúsculas. Así, los nombres NEWVAR, NewVar y newvar se consideran idénticos.

Respecto al Tipo de Variable se pueden elegir entre 8 tipos diferentes:

- **Numérico**. Una variable cuyos valores son números. Los valores se muestran en formato numérico estándar. El Editor de datos acepta valores numéricos en formato estándar o en notación científica.
- **Coma**. Una variable numérica cuyos valores se muestran con comas que delimitan cada tres posiciones y con el punto como delimitador decimal. El Editor de datos acepta valores numéricos para este tipo de variables con o sin comas, o bien en notación científica.
- **Punto**. Una variable numérica cuyos valores se muestran con puntos que delimitan cada tres posiciones y con la coma como delimitador decimal. El Editor de datos acepta valores numéricos para este tipo de variables con o sin puntos, o bien en notación científica.
- **Notación científica**. Una variable numérica cuyos valores se muestran con una E intercalada y un exponente con signo que representa una potencia de base diez. El Editor de Datos acepta para estas variables valores numéricos con o sin el exponente. El exponente puede aparecer precedido por una E o una D con un signo opcional, o bien sólo por el signo. Por ejemplo, 123, 1,23E2, 1,23D2, 1,23E+2 e incluso 1,23+2.
- **Fecha**. Una variable numérica cuyos valores se muestran en uno de los diferentes formatos de fecha_calendario y hora_reloj. Seleccione un formato de la lista. Puede introducir las fechas utilizando como delimitadores: barras, guiones, puntos, comas o espacios. El rango de siglo para los valores de año de dos dígitos está determinado por la configuración de las Opciones (menú Edición, Opciones, pestaña Datos).
- **Moneda personalizada**. Una variable numérica cuyos valores se muestran en uno de los formatos de moneda personalizados que se hayan definido previamente en la pestaña Moneda del cuadro de diálogo Opciones. Los caracteres definidos en la moneda personalizada no se pueden emplear en la introducción de datos pero sí se mostrarán en el Editor de Datos.
- **Cadena**. Variable cuyos valores no son numéricos y, por ello, no se utilizan en los cálculos. Pueden contener cualquier carácter siempre que no se exceda la longitud definida. Las mayúsculas y la minúsculas se consideran diferentes. También son conocidas como variables alfanuméricas.

Para definir el tipo se pulsa en la celda de intersección entre la variable y la columna, y una vez señalada la celda se pulsa en el icono que se muestra a la derecha  . Al pulsar este icono se muestra el cuadro con todos los tipos de variables como el que se muestra en la Figura 1.5.

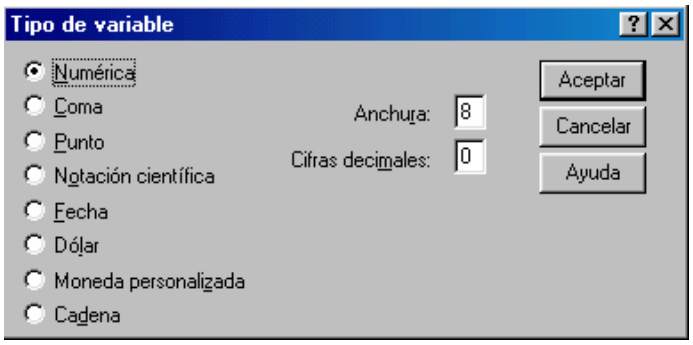


Figura 1.5 Cuadro de definición del tipo de variable

Las columnas designadas como **Anchura** y **Decimales**, se emplean para especificar la anchura y el número de decimales que contiene en las variables de tipo *Numérico*, *Coma*, *Punto*, *Notación científica*, *Dólar* y *Moneda personalizada*. Para las variables del tipo *Fecha*, se puede elegir entre un amplio abanico de formatos, y para las variables de tipo *Cadena* únicamente hay que especificar el número de caracteres máximos que tendrá dicha variable.

En la columna **Etiqueta**, se puede escribir un nombre para cada variable más descriptivo que el que proporcionan los 8 caracteres máximos del nombre de la variable.

En la columna **Valores**, se puede dar nombre a los valores numéricos de las variables nominales u ordinales. En el archivo *Datos de empleados* hay una serie de variables que son nominales (o categóricas), como por ejemplo *sexo*, *catlab* o *minoría*. Estas variables se han codificado numéricamente, pero los números asignados no tienen propiedades matemáticas, sino que representan categorías de las variables. Así *catlab* (categoría laboral), se ha codificado como 1, 2 ó 3, según el sujeto sea *Administrativo*, de *Seguridad*, o *Directivo*, respectivamente. Para asignar etiquetas a los valores, se pulsa en la celda correspondiente a la variable y se accede al cuadro que se muestra en la Figura 1.6.

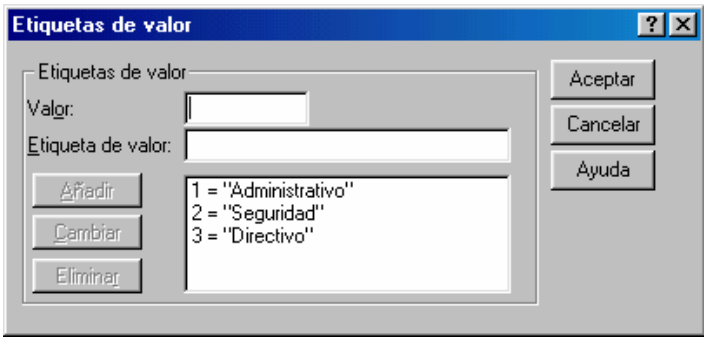


Figura 1.6. Cuadro para etiquetar los valores de variables nominales u ordinales

En muchas ocasiones no siempre se puede registrar para todos las variables todas las respuestas de los sujetos, bien porque el valor no se haya registrado o bien porque el sujeto se haya negado a contestar a alguna cuestión; estos casos no tiene validez de cara a los análisis y es preciso identificarlos de alguna manera. Una

opción que permite incluso identificar el origen de estos casos (si el registro se ha perdido, si el sujeto no sabe o no contesta, etc.) es la columna designada como **Perdidos**. Al pulsar en la celda correspondiente de la variable con dicha columna se activa a la derecha el mismo icono con los puntos suspensivos que al pulsar con el ratón nos lleva al cuadro que se muestra en la Figura 1.7.

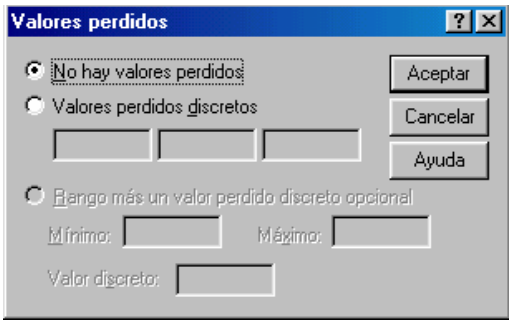


Figura 1.7 Cuadro para definir los valores perdidos

Se observa en este cuadro que se puede especificar como perdidos varios valores discretos, un rango de valores o un solo valor. El analista en cada caso determinará cuál de las opciones es más adecuada.

Por último, la columna **Alineación** permite definir en que posición de la celda (derecha, centro, izquierda) se visualiza el dato en el Editor de Datos. Y la columna **Escala**, permite determinar cómo es la variable: de escala (intervalo o razón), ordinal o nominal.

1.3 Definición y ejecución de un procedimiento

Para definir cualquier procedimiento de análisis estadístico, lo primero es disponer de datos en el Editor y, a continuación, elegir un procedimiento estadístico en la opción correspondiente del menú principal. En esta primera sesión confeccionaremos una distribución de frecuencias de la variable Categoría Laboral, del archivo *Datos de empleados*, para ello se sigue la secuencia:

Analizar – Estadísticos descriptivos – Frecuencias

y se muestra el cuadro de diálogo de la Figura 1.8.

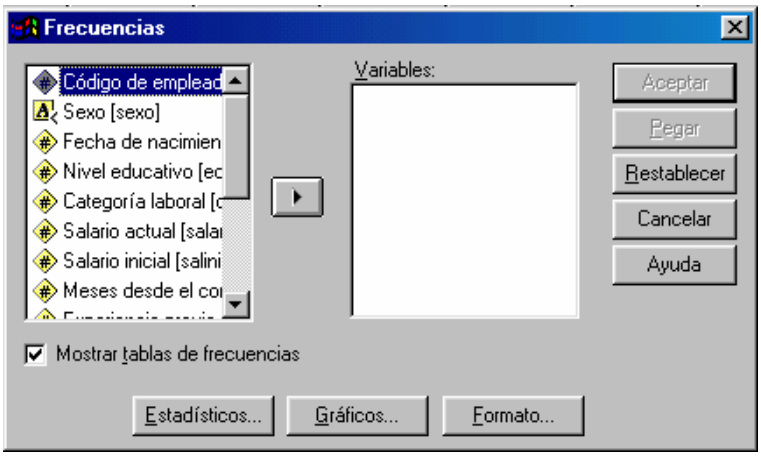


Fig. 1.8 Cuadro de diálogo del procedimiento Frecuencias

En la ventana de la izquierda se muestra la lista de variables que contiene el archivo de trabajo, de entre las cuales seleccionaremos la/s que se quiere/n analizar. Para realizar la selección, se marca cada variable con el puntero del ratón, y se traslada a la lista Variables, mediante la flecha intermedia. Cuando se han pasado las variables a la lista de variables se puede especificar los estadísticos descriptivos y los gráficos que se deseen, pulsando los botones correspondientes en la parte inferior del cuadro de diálogo. Los cuadros a los que se accede son los que se muestran en la Figura 1.9.

Fig. 1.9 Cuadros de diálogo de estadísticos (izquierda) y de gráficos del procedimiento Frecuencias

En el cuadro de estadísticos podemos señalar cualquiera de los que cuantifican los cuatro aspectos básicos de las distribuciones: los de posición (percentiles), los de tendencia central, los de variabilidad o dispersión y los de forma de la distribución (asimetría y curtosis). Como se ve en la Figura 1.9, por defecto no hay señalado ningún estadístico, y dado que la variable es categórica, tampoco lo vamos a requerir

Respecto a las opciones de gráficos, se puede elegir entre tres tipos, según sea el nivel de medida de la variable. Por defecto, la opción es no confeccionar ningún gráfico.

Pulsando, por último, el botón **Formato**, se puede elegir entre varios criterios de ordenación de la tabla de distribución, e incluso optar por no confeccionar

distribución alguna. Por defecto la opción es la de ordenación ascendente de valores

Una vez que está seleccionada la variable y señaladas todas las opciones, de estadísticos, de gráficos y de formato, pulsamos el botón **Aceptar** de la ventana de Frecuencias (Figura 1.7) y entramos en el interface de SPSS, denominado *Visor de SPSS*, cuya facilidad operativa es una de las varias características que lo distinguen favorablemente de las versiones 6 y anteriores.

1.4 Navegando por los Resultados

Como se ha dicho, cuando se pulsa el botón **Aceptar**, después de haber configurado las opciones del procedimiento requerido (en esta primera sesión una simple distribución de frecuencias, con su gráfico de pastel), el resultado se muestra en el *Visor*, cuyo aspecto se muestra en la Figura 1.10. La variable seleccionada para analizar es *Categoría laboral*, del archivo *Datos de empleados.*

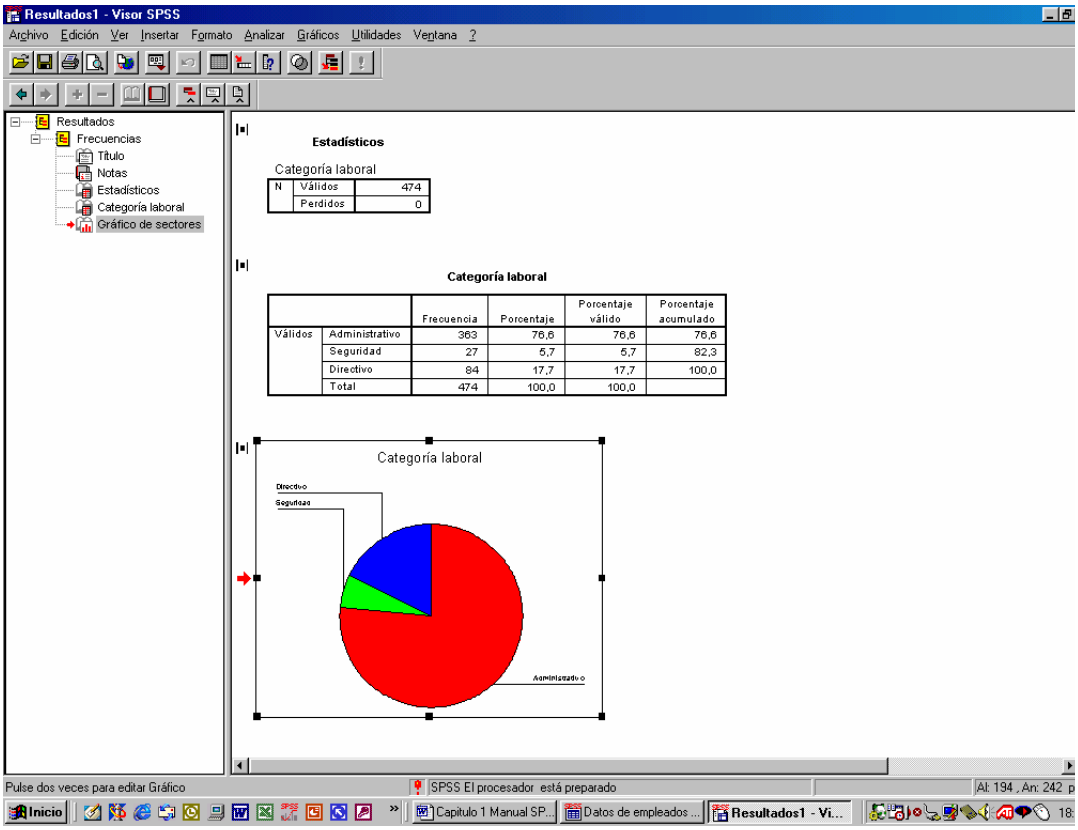


Fig. 1.10 *Visor de SPSS* con algunos resultados del procedimiento *Frecuencia*

Este interface consta de dos ventanas: la de la izquierda, con estructura de árbol, es, digamos, el guión o índice de los resultados que se muestran en la pantalla de la derecha. En el índice podemos señalar con el ratón cualquiera de los apartados, y verlo recuadrado en la ventana de la derecha. En la figura se puede ver señalado **Gráficos de sectores**, y en la de la derecha, el diagrama de sectores recuadrado y con una flecha de señal a la izquierda del recuadro. De este modo podemos navegar por los resultados con un simple clic del ratón en la parte que nos interese en cada momento.

1.5 Terminar una sesión con SPSS

Cuando ya se han cumplido los objetivos del análisis que hayamos podido efectuar con SPSS y se va a salir del programa, es conveniente guardar el trabajo realizado. Como ya se ha visto son varios los ámbitos en los que nos movemos en las sesiones de análisis, aunque sólo hemos visto dos de ellos: por un lado, el *Editor de datos*, y por otro, los resultados de los análisis que se muestran en el *Visor*. En el *Editor* se muestran los datos que hayamos leído, caso de que estuvieran almacenados en un archivo, o que hayamos escrito en el propio *Editor*. Respecto de los datos, sólo interesa archivarlos de nuevo cuando se ha efectuado alguna modificación de los mismos (recodificación de variables, creación de nuevas variables, etc.,); respecto de los resultados, el usuario determinará en cada momento si es conveniente su archivo para una posterior utilización.

2. Edición y transformación de datos

2.1 Edición de datos

Antes de proceder a introducir los datos en el Editor es necesario un trabajo previo, de lápiz y papel, para perfilar todo lo relativo a las variables: nombre de las variables, tipo de variables que se han registrado (numéricas, de cadena, de fecha, lógicas, etcétera), esquema de codificación de las variables, cuando éstas sean categóricas, u ordinales con pocos órdenes, especificación de los casos en que no se haya podido registrar el valor, y formato de presentación de las columnas que contienen las variables en el editor de datos.

2.1.1 Introducir datos en el Editor

La forma de entrar los datos en el editor es la misma que para cualquier hoja de cálculo. No obstante, antes de empezar a introducir los datos es conveniente definir las variables en la ventana de edición de variables, sobre todo en lo referente al tipo de variable, las etiquetas de los valores, los valores perdidos, y el formato de visualización en el editor. Una vez definidas las variables, en la ventana **Vista de datos** se comienza a teclear los valores. A diferencia de una hoja de cálculo, tipo ***Excel***, por ejemplo, es indistinto que después de ingresar cada dato se pulsa la tecla de <Retorno> o la tecla de <Tabulación>, pues en ambos casos se activa el caso inmediato inferior de la variable en la que se está tecleando los valores (recuerde el lector que en ***Excel***, si se pulsa el tabulador se pasa a la columna siguiente y si se pulsa retorno se pasa la fila siguiente). Si alguno de los datos se repite se puede utilizar los comandos de edición para abreviar la tarea.

2.1.2 Funciones de Edición

Con el Editor de datos se puede modificar un archivo de datos de varias maneras, a saber:

- Para *cambiar los valores de datos*, se pulsa en la casilla correspondiente al dato que se quiere reemplazar; este valor se muestra en el editor de casillas. Luego se introduce el nuevo valor y se pulsa <Retorno>.
- Para *cortar, copiar y pegar* se sitúa el cursor en la casilla que contiene el dato que se quiere cortar o copiar, y o bien se recurre a las teclas (Ctrl+X: corta; Ctrl+C: copia; Ctrl+V: pega), o bien se accede a estas funciones a través de Edición del menú principal, que despliega las opciones que se observan en la Figura 2.1

Fig. 2.1 Menú de Edición del Editor de datos. En la parte izquierda antes de haber copiado un elemento y en la derecha una vez que se ha copiado (se activa la opción Pegar)

La parte de la izquierda de esta figura tiene desactivada la opción de Pegar, y ello se debe a que todavía no se ha efectuado ninguna operación de cortado o copiado, mientras que en el menú de la derecha sí aparece la opción de pegar activada, después de haber realizado alguna de estas dos operaciones. Posteriormente, situamos el cursor en la celdilla en la que vayamos a pegar el dato cortado o copiado, y activamos la opción pegar.

- Para *añadir un nuevo caso* sólo hay que situarse en la primera celda de una fila vacía y teclear un dato. El editor inserta en el resto de las celdillas de esa fila (tantas como variables definidas) el valor perdido por el sistema. Si lo que se desea es insertar un caso entre los ya existentes, situamos el cursor debajo de la posición (caso o fila) donde queremos insertar el caso y en la opción de Datos del menú elegimos la opción Insertar caso.
- Para *insertar una nueva variable* se inserta un dato en una columna vacía y se crea automáticamente un nueva variable, con la definición por defecto, con todos los demás casos como valores perdidos por el sistema. Si lo que se quiere es insertar una variable entre otras que ya existen se procede igual que con la inserción de caso, pero en el sentido de las columnas o variables.
- Para *desplazar un variable* de sitio en el editor se marca la variable (pulsando el botón izquierdo del ratón sobre el nombre de variable) y se corta; luego se sitúa el cursor sobre el nombre de la variable en que quiere situarse la variable cortada, se inserta un nueva variable y, por último, se pega la variable cortada.

La definición de las variables se pueden cambiar en cualquier momento con sólo situar el ratón y pulsar en la cabecera de la variable, se accede a la rejilla de **Vista de variables**, donde se puede modificar cualquier aspecto de las variables.

Es frecuente que en un mismo archivo haya varias variables que, excepto el nombre, compartan las misma características; por ejemplo, las mismas etiquetas de respuesta, los mismos valores perdidos, etc. En ese caso no es preciso definir cada variable por separado, sino que se definen todos los aspectos (Tipo, Anchura, Decimales, Valores, Perdidos, etc.) para una de las variables y luego se copia y se pega en cada una de las variables que compartan esos mismos aspectos.

2.2 Creación de nuevas variables

SPSS permite crear nuevas variables a partir de las que ya existen en el archivo o bien las crea mediante las opciones de generación variables aleatorias que incorpora. En ambos casos el número de casos de las variables creadas es el mismo de los que hay en el archivo.

2.2.1 Creación de variables a partir de las que ya hay en el archivo

Para crear nuevas variables se pulsa:

Transformar – Calcular

y se accede al cuadro de dialogo de la Figura 2.2

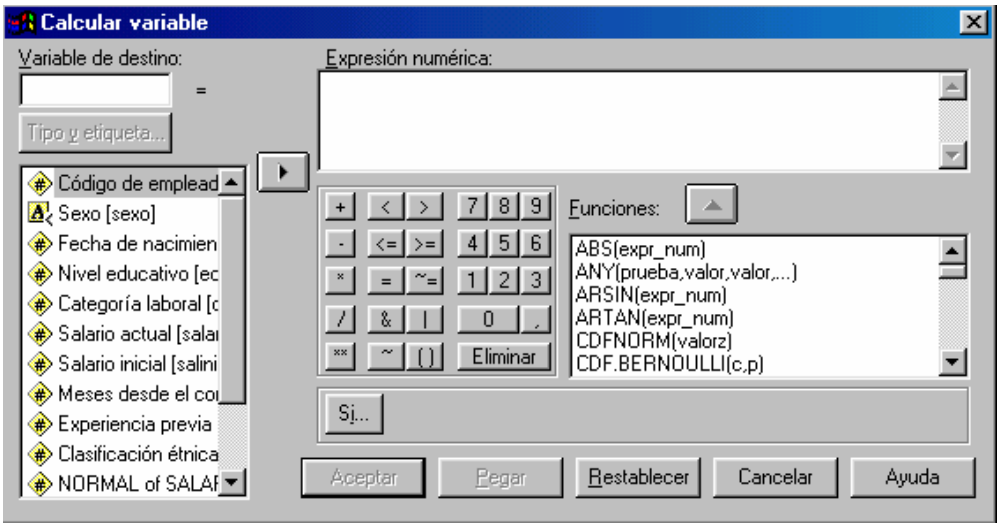


Figura 2.2 Cuadro de diálogo de creación de nuevas variables

En el cuadro **Variable de destino** se le da nombre a la nueva variable. En el momento que se teclea el primer carácter del nombre de la nueva variable se activa el botón Tipo y etiqueta, y se puede acceder a un cuadro en el que se define el tipo y se le da un nombre largo a la variable (el darle una etiqueta a la nueva variable es opcional; por defecto, las nuevas variables creadas son de tipo numérico con anchura 8 y 2 decimales). Una opción de etiqueta de variable es poner como tal la expresión numérica que va a servir para calcular la nueva variable. En el cuadro **Expresión numérica** se escribe la expresión que generará la nueva variable. Se puede observar que el cuadro de creación de variables incorpora un teclado con los operadores matemáticos, relacionales y lógicos comúnmente usados.

Como ejemplo, supongamos que se quiere crear una variable que nos indique el porcentaje de aumento que supone el salario actual respecto del salario inicial. En la Figura 2.3 se muestra la expresión numérica para el porcentaje

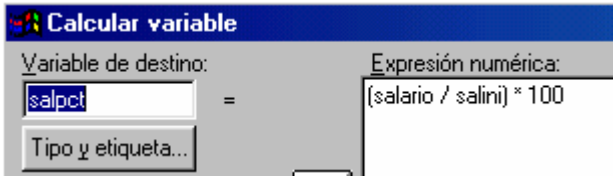


Figura 2.3 Expresión para obtener el porcentaje del salario actual respecto del inicial

Este procedimiento para crear variables es incondicional, es decir, la nueva variable tendrá valores en todos los casos, excepto en aquellos en los que alguna de las variables de la expresión numérica no tengan valor o sea un valor etiquetado como perdido. No obstante, es posible crear nuevas variables condicionada a valores de otras variables que haya en el archivo. para ello se pulsa el botón **Si...** y en el cuadro de diálogo (Figura 2.4) se establece la condición de creación de la nueva variable.

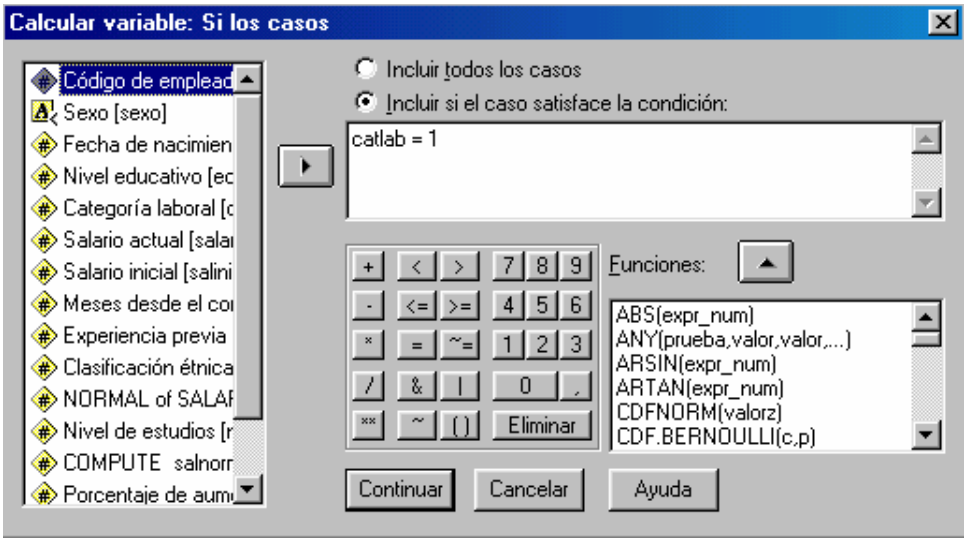


Figura 2.4 Cuadro para establecer la condición de creación de variables

En el caso que se muestra en la Figura 2.4, se ha establecido la condición de que la categoría laboral sea Administrativo (valor 1). En este caso, la nueva variable creada sólo tendrá valores en aquellos casos en que la variable categoría laboral tenga valor 1, mientras en el resto se mostrarán el signo de perdido del sistema (una coma en la celdilla).

2.2.3 Creación de una variable numérica a partir de una variable de fecha

En muchas ocasiones los archivos de datos contienen variables de tipo fecha que interesa convertirlos en una variable de tipo numérico para su inclusión en los análisis. Para esta conversión se emplean algunas funciones de conversión de fecha que incorpora SPSS. El inicio del tiempo en SPSS coincide con el año en que se instauró el calendario Gregoriano (1582), de tal modo que, por ejemplo, para convertir a días una variable de fecha hay que restar los días transcurridos desde la actualidad hasta 1582 de los días transcurridos de las fechas que contiene la variable de fecha. El archivo *Datos de empleado* contiene una variable de tipo fecha nombrada fechnac, cuyo formato es *Día*, *Mes*, *Año*, que se puede convertir en días mediante la expresión:

CTIME.DAYS(DATE.DMY(26,11,2001)-fechnac)

la función **CTIME.DAYS** convierte a días una expresión de fecha, mientras que la función **DATE.DMY** convierte a formato fecha el día, mes y año que se ponga en el paréntesis de la función. Una vez convertida una variable tipo fecha en días, se puede convertir en años dividiendo la expresión anterior por 365.252 (la parte decimal es para tener en cuenta los años bisiestos). Por último, con la función **TRUNC** se obtiene sólo la parte entera del resultado.

TRUNC(CTIME.DAYS(DATE.DMY(26,11,2001)-fechnac)/365.25)

2.2.4 Creación de variables aleatorias

Otra posibilidad de creación de variables es emplear las funciones de generación de variables aleatorias que dispone SPSS. Para ello, simplemente se da nombre a la nueva variable, se elige la función de probabilidad y se definen los parámetros de dicha función si es el caso. Las funciones de variable aleatoria que incorpora SPSS son las siguientes:

NORMAL(desv_típ) Numérico. Devuelve un número pseudo-aleatorio, distribuido normalmente, a partir de una distribución con media 0 y la desviación típica desv_típ, que debe ser un número positivo. Antes de cada generación, puede repetir la secuencia de números pseudo-aleatorios estableciendo la semilla en el cuadro de diálogo Semilla de aleatorización del menú Transformar.

RV.BERNOULLI(prob) Numérico. Devuelve un valor aleatorio de la distribución de Bernoulli, con el parámetro de probabilidad prob especificado.

RV.BETA(forma1,forma2) Numérico. Devuelve un valor aleatorio de una distribución Beta, con los parámetros de forma especificados.

RV.BINOM(n,prob) Numérico. Devuelve un valor aleatorio de la distribución Binomial, con el número de intentos y el parámetro de probabilidad especificados.

RV.CAUCHY(loc,escala) Numérico. Devuelve un valor aleatorio de la distribución de Cauchy, con los parámetros de posición y escala especificados.

RV.CHISQ(gl) Numérico. Devuelve un valor aleatorio de la distribución de chi-cuadrado, con los grados de libertad gl especificados.

RV.EXP(forma) Numérico. Devuelve un valor aleatorio de una distribución exponencial, con el parámetro de forma especificado.

RV.F(gl1,gl2) Numérico. Devuelve un valor aleatorio de la distribución F, con los grados de libertad gl1 y gl2 especificados.

RV.GAMMA(forma,escala) Numérico. Devuelve un valor aleatorio de la distribución Gamma, con los parámetros de forma y escala especificados.

2 Para el cálculo de nuevas variables, en las expresiones numéricas los decimales se escriben con punto

RV.GEOM(prob) Numérico. Devuelve un valor aleatorio de una distribución Geométrica, con el parámetro de probabilidad especificado.

RV.HYPER(total,muestra,aciertos) Numérico. Devuelve un valor aleatorio de la distribución Hipergeométrica, con los parámetros especificados.

RV.LAPLACE(media,escala) Numérico. Devuelve un valor aleatorio de la distribución de Laplace, con los parámetros de media y escala especificados.

RV.LOGISTIC(media,escala) Numérico. Devuelve un valor aleatorio de la distribución Logística, con los parámetros de media y escala especificados.

RV.LNORMAL(a,b) Numérico. Devuelve un valor aleatorio de la distribución log-normal, con los parámetros especificados.

RV.NEGBIN(umbral,prob) Numérico. Devuelve un valor aleatorio de la distribución Binomial negativa, con los parámetros de umbral y probabilidad especificados.

RV.NORMAL(media,desv_típ) Numérico. Devuelve un valor aleatorio de la distribución normal, con la media y la desviación típica especificadas.

RV.PARETO(umbral,forma) Numérico. Devuelve un valor aleatorio de la distribución de Pareto, con los parámetros de umbral y forma especificados.

RV.POISSON(media) Numérico. Devuelve un valor aleatorio de la distribución de Poisson, con el parámetros de media o tasa especificado.

RV.T(gl) Numérico. Devuelve un valor aleatorio de la distribución t de Student, con los grados de libertad gl especificados.

RV.UNIFORM(mín,máx) Numérico. Devuelve un valor aleatorio de la distribución uniforme, con el mínimo y el máximo especificados. Véase también la función UNIFORM.

RV.WEIBULL(a,b) Numérico. Devuelve un valor aleatorio de la distribución de Weibull, con los parámetros especificados.

UNIFORM(máx) Numérico. Devuelve un número pseudo-aleatorio distribuido uniformemente entre 0 y el argumento máx, el cual debe ser numérico (pero puede ser negativo). Puede repetir la secuencia de números pseudo-aleatorios estableciendo la misma semilla de aleatorización (disponible en el menú Transformar) antes de cada generación.

Otro tipo de funciones de SPSS, que el lector puede encontrar en la ayuda del programa (pulsando F1 se accede a la ayuda) son las siguientes:

- Funciones aritméticas
- Funciones estadísticas
- Funciones de cadena
- Funciones de fecha y hora
- Funciones de distribución
- Funciones de variables aleatorias
- Funciones de valores perdidos

2.3 Recodificación de variables

En ocasiones interesa hacer una aproximación inicial a los datos, de modo que sea preciso realizar una recodificación, como por ejemplo, convertir una variable cuantitativa en cualitativa. Son varias las formas de recodificación:

- Recodificar en las mismas variables
- Recodificar en distintas variables
- Recodificación automática

Mediante la primera opción se recodifica los valores de una variable, y ésta pierde sus valores originales por los valores que resulten de la codificación. Sin embargo, esta forma de recodificación tiene el inconveniente de que se pierde los datos originales de esa variable. Por esta razón, sólo es recomendable cuando haya seguridad de que los datos originales no se van a necesitar en adelante. Para recodificar en distintas variables se sigue la secuencia,

Transformar → Recodificar → En distintas variables

y se accede al cuadro de diálogo que se presenta en la Figura 2.5. En ese cuadro se elige la variable que se quiere recodificar y se incorpora a la lista **Var. numérica →Var. de resultado**. En los campos **Nombre** y **Etiqueta** se sitúa el nombre de la nueva variable y, si se quiere, la etiqueta o descripción de la nueva variable. Nombrada la variable se procede a recodificar pulsando el botón **Valores antiguos y nuevos** y se accede al cuadro de la Figura 2.6. Hay varias posibilidades de recodificación: desde valores discretos a rangos de valores, recodificación de valores perdidos, etcétera.

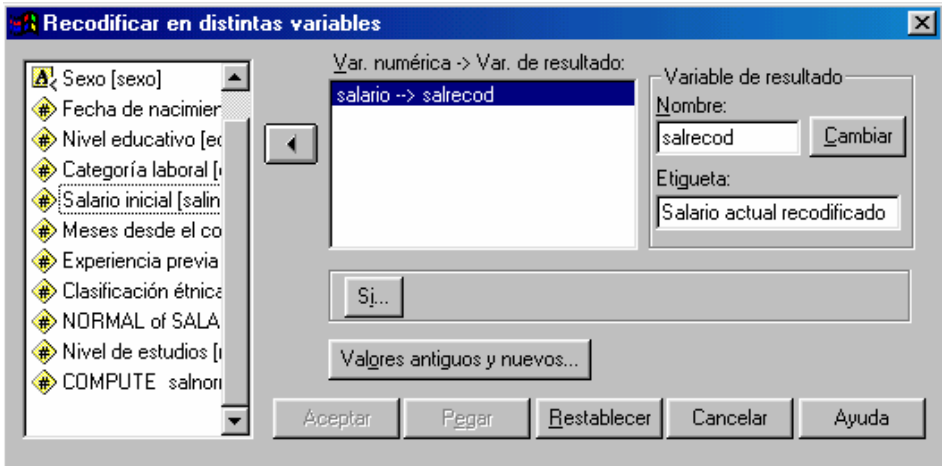


Figura 2.5 Cuadro de selección de variables a recodificar

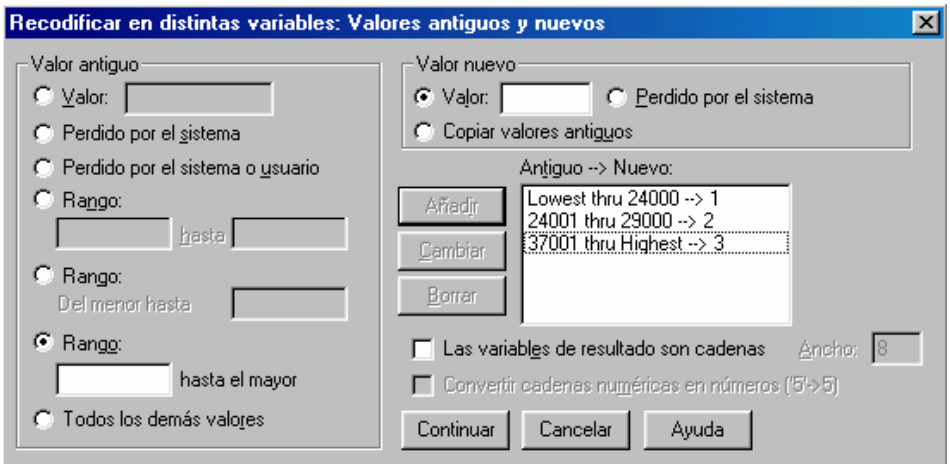


Figura 2.6 Cuadro para establecer los valores de recodificación

Como ejemplo se ha recodificado el salario actual (variable salario del archivo *Datos de empleados*), de tal modo que se ha establecido tres categorías para recodificar el salario, asignando, en la nueva variable, el valor 1 a los salarios iguales o inferiores a 24000$, el valor 2 a los salarios entre 24001$ y 29000$, y el valor 3 a los salarios por encima de 29000$. Una vez que se ha determinado el valor o rango de valores a recodificar y el nuevo valor se pulsa el botón **Añadir** y se incorpora a la lista **Antiguo→ Nuevo**. Las entradas en esta lista se pueden cambiar o borrar, marcando las entradas correspondientes y pulsando el botón que interese.

Al igual que en el proceso de creación de variables, también se pueden recodificar variables condicionada a los valores de otra/s variable/s del archivo. Para establecer la condición hay que pulsar el botín **Si...** del cuadro de la Figura 2.5 y se muestra el mismo cuadro para establecer las condiciones ya visto en la Figura 2.4.

2.4 Recodificación automática

Algunos de los procedimientos del SPSS sólo permiten variables de tipo numérico. Sin embargo, en muchas ocasiones los archivos contienen variables de cadena que es preciso someter al SPSS, por ejemplo, para construir una tabla con información resumida, y para ello es necesario previamente transformar dicha variable de cadena en una variable de tipo numérico, pero sin que se pierda la información que la variable contiene. Para efectuar esta recodificación, SPSS dispone de un procedimiento mediante el cual una variable de tipo cadena la recodifica siguiendo un orden alfabético en una variable numérica, y a cada valor numérico resultante le asigna como etiqueta el nombre que contiene la variable en cada caso. La secuencia será:

Transformar → Recodificación automática

Como ejemplo, supongamos que en uno de nuestros archivos una de las variables contiene el nombre de una serie de colegios en los que estamos llevando a cabo un investigación determinada. Los nombres de los colegios los habremos introducido en una variable de tipo cadena, pero después necesitaremos convertir

esta variable a otra de tipo numérico. Los nombres de los colegios se muestran en la parte izquierda de la Figura 2.7, mientras que en la derecha se muestra el cuadro de diálogo de recodificación automática.

Figura 2.7 Variable de tipo cadena y cuadro de diálogo de recodificación automática

En este cuadro de diálogo se selecciona la variable que se quiere recodificar y se incorpora a la lista Variable -> Nuevo nombre. En el campo adyacente al botón Nuevo nombre se da nombre a la variable de salida y una vez escrito se pulsa el botón y se incorpora a la lista. Después de aceptar el procedimiento, en el Visor de resultados se muestra un cuadro de texto que informa de la recodificación y de cuáles son los valores numéricos de los registros de la variable de tipo cadena. El cuadro de texto para los diez casos de colegios será el siguiente:

COLEGIO	NCOLEGIO
Old Value	New Value Value Label
Antonio Machado	1 Antonio Machado
Antonio Salinas	2 Antonio Salinas
Cesar Vallejo	3 Cesar Vallejo
Federico G. Lorca	4 Federico G. Lorca
Gabriel Celaya	5 Gabriel Celaya
J.L. Borges	6 J.L. Borges
León Felipe	7 León Felipe
Luis Panero	8 Luis Panero
Miguel Hernández	9 Miguel Hernández
Pablo Neruda	10 Pablo Neruda

La nueva variable se crea a partir de un orden alfabético ascendente o descendente (según se especifica en la opción correspondiente del cuadro de diálogo) y es de tipo numérico, y asigna como etiqueta (Value Label) el nombre correspondiente.

2.5 Asignación de rangos a casos

Otra opción de SPSS es la de asignar rangos a casos es decir, ordenar una variable según un orden ascendente o descendente de los valores y asignarlos un número de orden. A la variable de salida no es preciso darle un nombre, pues el propio programa lo hace antecediendo la letra r al nombre de la variable que se ha ordenado. La secuencia para asignar rangos y acceder al cuadro de diálogo de la Figura 2.8, es la siguiente:

Transformar → Asignar rangos a casos...

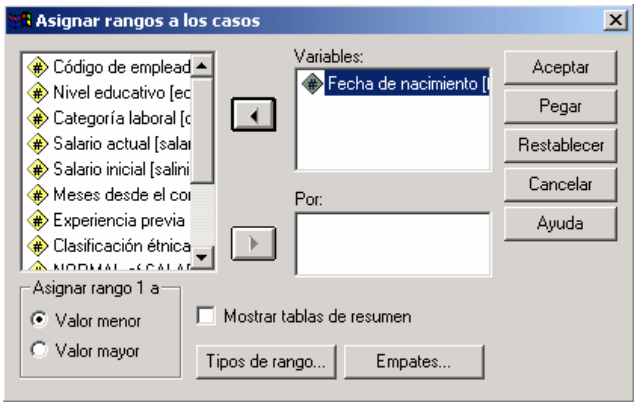


Figura 2.8 Cuadro de diálogo para asignar rangos a casos

Uno de los aspectos que hay que considerar es el de los empates de valores y decidir el criterio de asignación de rangos, para ello se pulsa el botón correspondiente a **Empates** y se muestra el cuadro de la Figura 2.9(a). Se puede elegir entre asignar el rango medio el menor o el mayor o bien asignar tantos rangos cómo valores distintos haya.

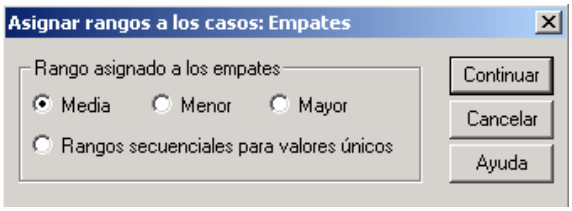


Figura 2.9(a)

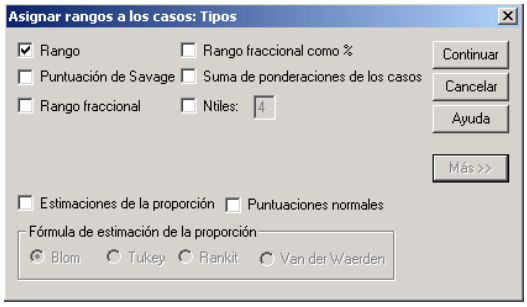
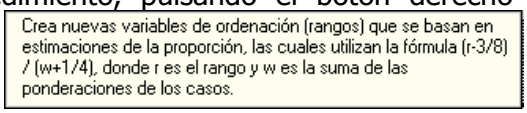


Figura 2.9(b)

Figuras 2.9 (a) Tratamiento de empates en la asignación de rangos a casos y (b) Tipos de rangos en la asignación de rangos a casos

Además de los empates, también se puede establecer el tipo de rango, e incluso normalizar las puntuaciones. Para ello se pulsa el botón **Tipo de rango** y se muestra el cuadro de la Figura 2.9(b). Por defecto el tipo es el de rango simple, pero hay varios opciones más (Puntuación de Savage; Rango fraccional; etc.) cuyo significado puede el lector consultar situando el puntero del ratón sobre el nombre de dicha opción y pulsar el botón derecho de modo que en pantalla aparece un cuadro blanco con la explicación correspondiente. Por ejemplo, si deseamos

normalizar las puntuaciones mediante el procedimiento de Blom y deseamos saber cuál es el procedimiento, pulsando el botón derecho del ratón obtenemos el siguiente cuadro:



2.6 Contar apariciones de casos

En determinadas situaciones de análisis es preciso contar el número de veces que los sujetos responden un valor o grupo de valores determinados. Piense el lector por ejemplo en las respuestas a un test con un determinado número de alternativas por item. Para ello se sigue la secuencia:

Transformar → Contar apariciones...

y se accede al cuadro de diálogo de la Figura 2.10(a). Una vez nombrada la variable destino y seleccionadas las variables sobre las que se va a establecer el conteo, se pulsa el botón Definir valores y se accede al cuadro de la Figura 2.10(b), donde se escribe el valor o rango de valores en la parte izquierda de dicho cuadro y se añaden, mediante el botón Añadir a la ventana Contar los valores

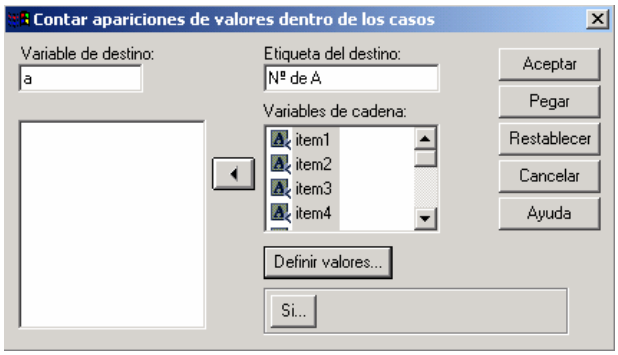


Figura 2.10(a)

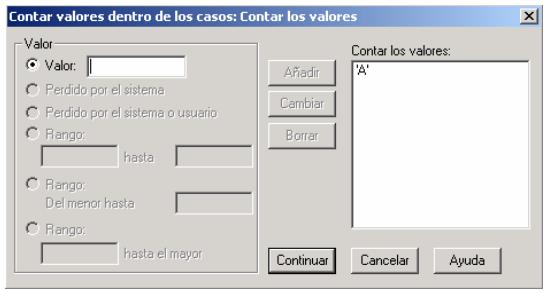


Figura 2.10(b)

Figuras 2.10 (a) Cuadro de diálogo para seleccionar variables sobre las que contar valores y (b) Cuadro para determinar los valores o rango de valores a contar

Como ejemplo, contamos el valor A para los siguientes 10 ítems en un conjunto de 5 casos. La nueva variables creada a, de tipo numérico, contiene el número de veces que cada sujeto contesta la alternativa A en los diez ítems de la prueba.

	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10	a
1	A	B	B	C	A	D	D	A	C	D	3,00
2	B	C	A	C	D	D	A	D	C	D	2,00
3	C	D	A	D	D	C	B	B	C	A	2,00
4	D	D	A	A	B	B	C	C	B	B	2,00
5	C	A	C	B	B	A	D	C	B	C	2,00

Al igual que en muchos de los procedimientos vistos en este tema, también se puede determinar un conteo de valores, condicionado a algún valor o valores de las variables que contenga el archivo de datos. Para establecer la condición se pulsa el

botón Si... y se accede al cuadro de diálogo, ya visto en la Figura 2.4, para establecer la condición para el conteo.

3. Manipulación de archivos

3.1 Introducción

En la mayoría de los procesos de análisis es preciso organizar el archivo de trabajo de alguna manera determinada. En algunos momentos tendremos que ordenarlo de acuerdo a alguna o algunas de las variables; en otros, deberemos seleccionar sólo un conjunto de casos para efectuar análisis sobre dicho conjunto. En otras ocasiones, interesará proceder a generar variables que resuman algunas de las variables del archivo y guardar dicha información en otro archivo para un uso posterior. O también sucederá que los datos los tengamos repartidos entre varios archivos, de modo que, previo al análisis, será preciso fusionarlos. En este capítulo, aprenderemos a efectuar estas y otras operaciones, las cuales se encuentran en la opción Datos del menú principal.

3.2 Ordenar casos

Esta opción permite ordenar el archivo de acuerdo a una o más variables en sentido ascendente o descendente (por defecto, el primero). Para la ordenación por dos o más variables se ordena según la primera variable especificada y la ordenación para la segunda se realizará dentro de cada uno de los valores de la primera, y así sucesivamente. Para acceder al procedimiento:

Datos → Ordenar casos...

y se muestra al cuadro que de la Figura 3.1, en el cual se selecciona/n la/s variable/s por la/s que se va a ordenar el archivo.

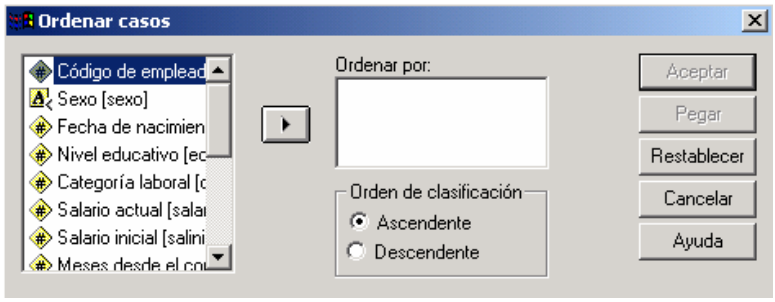


Figura 3.1 Cuadro de diálogo para ordenar casos

Como ejemplo de ordenación ascendente se puede ver las variables v1 y v2 antes y después de ordenadas, primero en v1 y, anidada, en v2

	v1	v2
1	3	3
2	1	4
3	2	5
4	6	3
5	1	6
6	3	8
7	2	9
8	3	10
9	2	12
10	1	7
11	3	4
12	4	3
13	3	2
14	3	5
15	5	1

	v1	v2
1	1	4
2	1	6
3	1	7
4	2	5
5	2	9
6	2	12
7	3	2
8	3	3
9	3	4
10	3	5
11	3	8
12	3	10
13	4	3
14	5	1
15	6	3

3.3 Selección de casos

Los procesos de análisis se pueden efectuar sobre el total de datos que hay en un archivo, o sobre un subconjunto de datos. SPSS ofrece varios métodos para seleccionar conjuntos de datos, pero básicamente son tres los criterios que se pueden seguir a la hora de seleccionar casos:

- Selección en función de valores de variables
- Selección de una muestra aleatoria de casos
- Selección de un rango determinado de casos

Para acceder a la selección de casos se sigue la secuencia:

Datos →Selección de casos...

y se muestra el cuadro de diálogo de la Figura 3.2

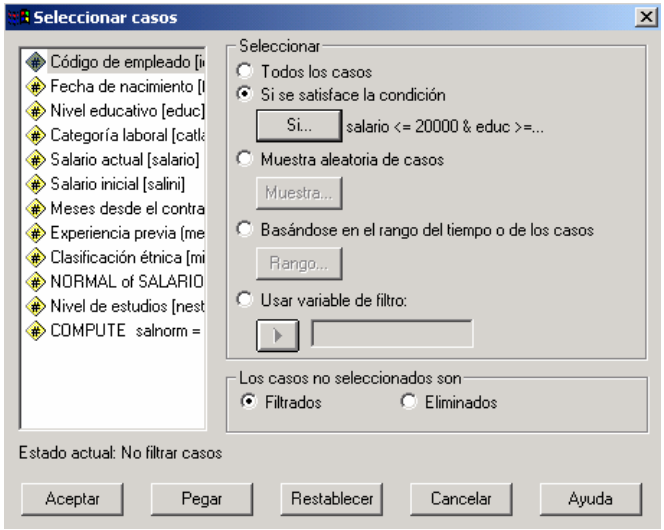


Figura 3.2 Cuadro de opciones para seleccionar casos

Por defecto, esta activada la opción de utilizar todos los casos. Una vez establecido el criterio de selección, se debe determinar si la selección será temporal o permanente, y para ello se señala la opción correspondiente en el recuadro **Los**

casos no seleccionados son. Si se señala la opción **Filtrados** (por defecto) los procedimientos de análisis sólo tomarán en consideración los casos seleccionados, mientras los no seleccionados se muestran con una señal (/) en el editor de datos. Si se señala la opción **Eliminados**, los casos no seleccionados son eliminados del archivo de trabajo, razón por la cual, si se quiere utilizar para posteriores análisis, será preciso volver a leer el archivo que los contiene. El lector puede colegir que la opción de eliminar los casos no seleccionados sólo se debe utilizar cuando efectivamente no se vayan a emplear más estos casos, y lo más prudente es simplemente filtrarlos.

Siempre que se efectúa un proceso de selección SPSS crea automáticamente un variable denominada *filter_$*, con dos únicos valores, 0 y 1, etiquetados como *No seleccionados* y *Seleccionados*, respectivamente. Esta variable se puede cambiar de nombre y utilizar en un proceso de selección posterior, incorporándola al campo **Usar variable de filtro**. Hay que advertir al lector, que si no se renombra la variable de filtro creada, cada vez que se realiza una nueva selección la variable de filtro es reemplazada por una nueva con el mismo nombre, y por tanto se pierde la memoria de los casos que fueron seleccionados en el proceso de selección anterior.

3.3.1 Selección en función de valores de variables

Este modo de selección sigue las mismas pautas que ya se han explicado cuando se crean o recodifican variables de acuerdo a una o varias condiciones. Para acceder al cuadro de selección condicional se pulsa el botón **Si...**, y se escribe la condición. Como ejemplo, en el archivo *Datos de empleados* se ha realizado una selección de aquellos casos cuyo salario es inferior o igual a 20000 dólares y han estudiado 10 años o más. De acuerdo a este criterio el número de casos seleccionados han sido 22, 1 hombre y 21 mujeres.

3.3.2 Selección de una muestra aleatoria de casos

Esta opción de selección es muy útil cuando se quieren construir, por ejemplo, modelos de regresión sobre sólo un conjunto de casos, y posteriormente comprobar si dicho modelo es extrapolable a otros conjuntos del total de casos que componen el archivo de datos. Para acceder a este tipo de selección aleatoria, se señala la opción correspondiente y se pulsa el botón **Muestra**, mostrándose el cuadro de la Figura 3.3

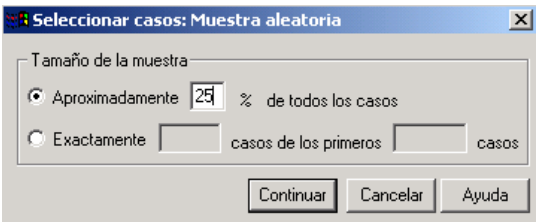


Figura 3.3 Cuadro de selección aleatoria de casos

Se puede elegir en términos de porcentaje o bien especificar un cantidad de casos de los primeros n casos. En ambos casos, SPSS emplea una semilla de aleatorización diferente para cada proceso, aunque es posible establecer una misma semilla para todos los procesos, cuyo resultado sería que las muestras

contendrían siempre los mismos casos. La opción para establecer la semilla se encuentra en el menú Transformar.

3.3.3 Selección según un rango de tiempo o de casos

Para realizar una selección basándose en un rango de tiempo es preciso previamente haber definido alguna variable de fecha, que es una opción de Datos en el menú principal (sugerimos al lector que explore esta posibilidad de definir variables de fecha). Si se han definido este tipo de variables, sólo es posible establecer una rango en base a estas variables de fecha. Si no se ha definido este tipo de variable sólo se puede seleccionar un rango de acuerdo a la situación de los casos. El cuadro para determinar el rango según los casos se muestra en la Figura 3.4

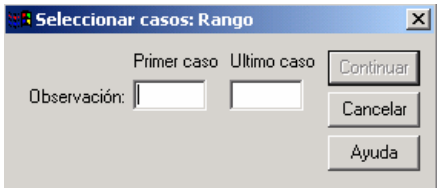


Figura 3.4 Cuadro para seleccionar un rango de casos

3.4 Agregación de datos

Cuando un archivo contiene variables de agrupamiento, es posible extraer información resumen de otras variables en función de los valores o categorias de las variables de agrupamiento, y construir un nuevo archivo con esta información estadística. El archivo así construido, contendrá tantos casos como categorías tenga la variable de agrupamiento y tantas variables como se creen más la propia variable de agrupamiento. Si se emplean varias variables de agrupamiento, el número de casos del nuevo archivo será igual al producto del número de categorías de cada una de las variables de agrupamiento empleadas. Si, por ejemplo, se emplearan tres variables de agrupamiento, la primera con dos categoría, la segunda con cuatro y la tercera con tres, el total de casos del archivo con información resumen será de 2x4x3 = 24 casos.

Para ilustrar el procedimiento utilizaremos el archivo *Datos de empleados* que contiene varias variables de agrupamiento, y otras de escala que puede servir para extraer información resumida. Las variables de agrupamiento son **sexo**, **categoría laboral** y **minoría**. Para acceder al cuadro de diálogo que se muestra en la Figura 3.5 se sigue la secuencia

Datos → Agregar...

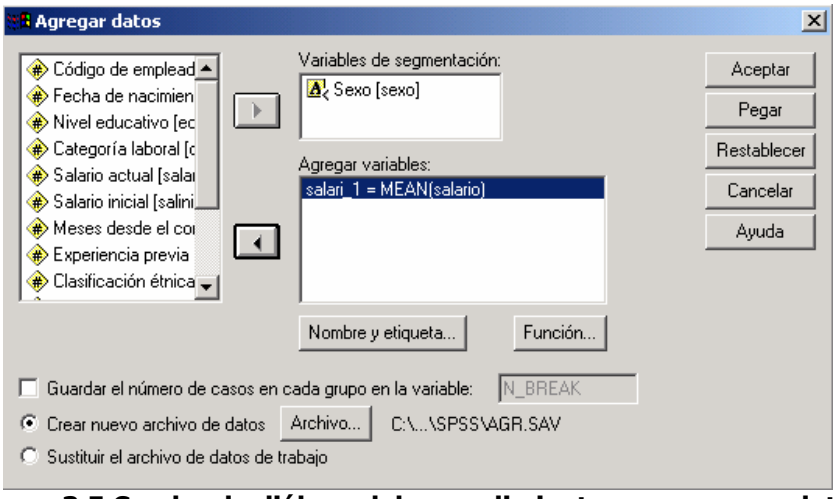


Figura 3.5 Cuadro de diálogo del procedimiento para agregar datos

A la lista **Variables de segmentación** se incorpora la variable o variables de agrupamiento y a la lista **Agregar variables** se incorporan la variable o variables de las que queremos extraer información resumida. Observará el lector, que las variables que se incorporan a la lista **Variables de segmentación**, desaparecen de la lista de variables de la ventana izquierda del cuadro, mientras que las variables que se incorporan a la lista **Agregar variables**, permanecen en el listado general de variables. La razón es obvia, ya que sobre una misma variable se puede obtener más de un estadístico, y por tanto se puede elegir la misma variable varias veces.

Una vez elegida la variable se pasa a la lista **Agregar variables** y, de manera automática, se añade un guión bajo y un 1 a la raíz del nombre de la variable elegida, y por defecto elige como función agregada la Media (MEAN). Si eligiéramos la misma variable de nuevo se añadiría un guión bajo y un 2 a dicha variable y así sucesivamente. No obstante esta manera automática de renombrar la variable de salida, se puede cambiar tanto el nombre como la función agregada que se quiere obtener. Para cambiar el nombre, se pulsa en el botón **Nombre y etiqueta**, y se accede al cuadro que se muestra en la Figura 3.6 (a) y para cambiar la función estadística se pulsa en el botón **Función** y se muestra la Figura 3.6 (b).

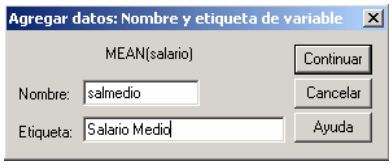


Figura 3.6 (a)

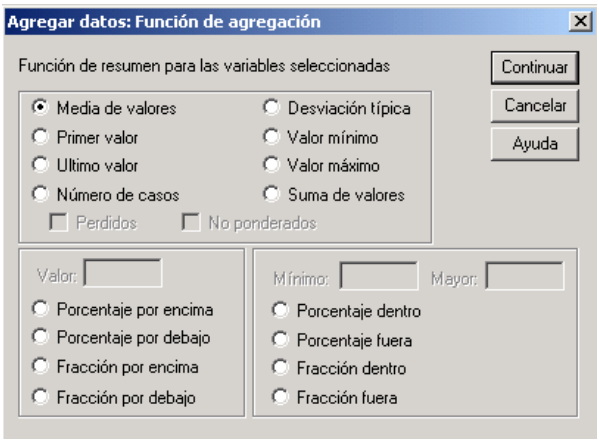


Figura 3.6 (b)

Figuras 3.6 (a) Cuadro para cambiar el nombre y etiqueta de la variable de salida y (b) cuadro para elegir la función de agregación.

Además de elegir las variables de segmentación y la agregadas, se puede dar nombre al archivo generado, aunque por defecto se nombra, si no se cambia, como *AGR.SAV*. El lector debe saber que si no se cambia el nombre del archivo de salida, cualquier nuevo procedimiento de agregación sobrescribirá el archivo anteriormente creado. Por último, se puede optar porque el archivo creado sea el nuevo archivo de trabajo, señalando dicha opción en la parte inferior del cuadro.

El archivo generado, como ya se ha dicho, tendrá dos variables, la de agrupamiento y la del salario promedio, y dos casos, tantos como categorías de la variable de agrupamiento.

	sexo	salmedio
1	Ho	41441,78
2	Mu	26031,92
3		
4		

Por defecto, la variables numéricas de salida son del tipo numérico y anchura ocho con dos decimales. Si se quiere cambiar el tipo, se procederá de la manera descrita en el capítulo 1.

Si se utiliza más de una variable de agrupamiento y se pide más de una función agregada, el aspecto del cuadro de diálogo es el que se muestra en la Figura 3.7. Además de las variables utilizadas, se ha especificado que el archivo generado sea el nuevo archivo de trabajo y que se genere una nueva variable con el número de casos para cada combinación de las categorías de las variables de segmentación. Dado que las categorías de las variables de segmentación, **sexo**, **categoría laboral** y **clasificación étnica**, son 2, 3 y 2, respectivamente, el número de casos del archivo generado serán 12 y el las variables serán las tres de agrupamiento más las cuatro con información agregada más la variable con el número de casos, en total 8 variables.

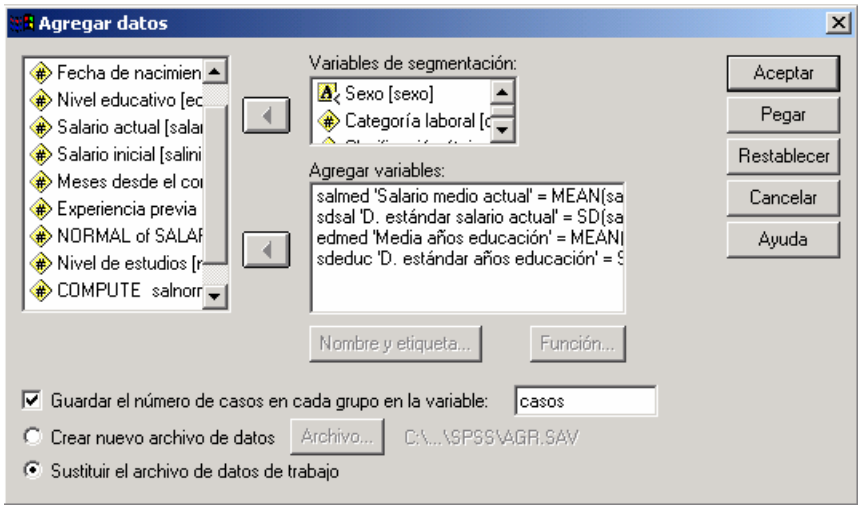


Figura 3.7 Cuadro de agregación de datos con varias variables de segmentación y varias variables agregadas.

En el cuadro inferior se puede ver el contenido del archivo resultante, en el cual se observa que sólo hay 9 casos y no los 12 pronosticados, y ello es debido a que no hay mujeres directivas de raza minoritaria, ni hay mujeres empleadas en Seguridad. A este archivo que contiene información agregada lo hemos guardado con el nombre *Datos agregados según categoría laboral*, y nos servirá para ilustrar algunos aspectos del procedimiento para fusionar archivos

	sexo	catlab	minoría	salmed	sdsal	edmed	sdeduc	casos
1	Ho	Administ	No	32671,64	8579,00	13,87	2,05	110
2	Ho	Administ	Sí	28952,13	5712,42	13,47	2,42	47
3	Ho	Segurida	No	31178,57	1658,74	10,29	2,05	14
4	Ho	Segurida	Sí	30680,77	2562,92	10,08	2,47	13
5	Ho	Directivo	No	65683,57	18029,45	17,50	1,54	70
6	Ho	Directivo	Sí	76037,50	17821,96	16,00	2,94	4
7	Mu	Administ	No	25471,45	6092,37	12,12	2,30	166
8	Mu	Administ	Sí	23062,50	3972,37	12,50	1,91	40
9	Mu	Directivo	No	47213,50	8501,25	16,00	,00	10

3.6 Fusión de archivos

En muchas ocasiones, los datos relativos a un mismo proyecto de trabajo suelen estar repartidos en diferentes archivos, y para el análisis de datos es preciso fusionar estos archivos en uno sólo. Hay dos posibilidades de fusión:

- Añadir casos. Los archivos contienen las mismas variables pero casos diferentes.
- Añadir variables. los archivos contienen los mismos casos pero diferentes variables.

Para ilustrar ambos procedimientos se ha dividido el archivo *Datos de empleados* en varios archivos. En primer lugar, el archivo se ha partido en dos archivos, uno conteniendo los casos 1 a 220 (previamente el archivo se ha ordenado por la variable **id –código de empleado-**) y lo hemos guardado con el nombre *Datos de*

empleados 1 – 220, y el otro, con los casos 221 a 474, lo hemos guardado con el nombre *Datos de empleados 221 – 474*. En el primer archivo, además, se ha modificado el nombre de la variable **fechnac** por **nacim**.

En segundo lugar, el archivo *Datos de empleados* se ha partido en dos. El primero contiene las variables **id**, **sexo**, **fechnac**, **educ**, **catlab** y **salini** y lo hemos guardado con el nombre *Datos de empleados con salario inicial*, y el segundo contiene las variables **id**, **salario**, **tiempemp**, **expprev** y **minoría**, y lohemos guardado con el nombre *Datos de empleados con salario actual*.

3.6.1 Añadir casos

Lo primero es tener como archivo de trabajo alguno de los archivos que vamos a fusionar. El orden de los archivos a fusionar es irrelevante pues siempre se puede, una vez fusionados, ordenar los casos según la/s variable/s que queramos. En este caso vamos a abrir el archivo *Datos de empleados 1 – 220*. Una vez abierto se tiene que seleccionar el archivo con el que lo vamos a fundir. Para ello se pulsa:

Datos → Fundir archivos → Añadir casos

y se accede al cuadro de diálogo de la Figura 3.8.

Figura 3.8. Cuadro de diálogo de Añadir casos: Leer archivo

En el cuadro se muestran todos los archivos del directorio de datos por defecto. Se marca el archivo externo, *Datos de empleados 221 – 474*, que vamos a fusionar con el que ya está activo, y luego se pulsa el botón **Abrir**. Entonces se muestra el cuadro de diálogo que aparece en la Figura 3.9. Si el nombre de las variables en el archivo de trabajo y en el archivo externo son iguales, en la lista **Variables en el nuevo archivo de datos de trabajo**, se muestran las variables que tendrá el archivo resultante de la fusión, que llamaremos archivo combinado. Si, como es el caso, el nombre de alguna variable difiere en uno y otro archivo, se muestran en la lista **Variables desemparejadas**. La variable seguida de un asterisco es la variable del archivo de trabajo, y la variable seguida del signo más es la variable que aporta el archivo externo. El que haya variables desemparejadas puede deberse a alguna de estas circunstancias:

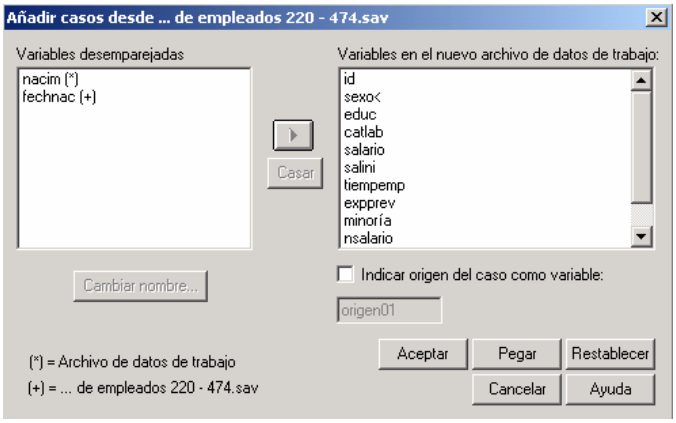


Figura 3.9 Cuadro de diálogo Añadir casos desde...

- Variables que se encuentran en un archivo sólo (es nuestro caso, aunque el contenido de las variables en uno y otro archivo es el mismo: casos de una misma variable, los nombres son diferentes)
- Variables definidas como numéricas en un archivo y como de cadena en el otro, lo cual es de imposible combinación.
- Variables que aun siendo ambas de cadena, el ancho sea diferente en uno y otro archivo.

En el caso de variables desemparejadas, en el que las dos contienen información sobre la misma variable, lo habitual es cambiar de nombre a una de las variables y nombrarla como la otra, luego marcar ambas variables, lo que activa el botón **Casar**, pulsar la flecha de arriba y pasarla a la lista **Variables en el nuevo archivo...**

La otra opción es marcar ambas variables, sin cambiar el nombre, con lo que se activa el botón Casar, pulsar la flecha de arriba de pasarla al cuadro de la Variables en el nuevo archivo... El nombre de la variable en el archivo combinado será el mismo que el del archivo de trabajo, aunque en la lista de variables del nuevo archivo aparezca como nacim&fechnac.

Por último, siempre es posible pasar al cuadro de la lista de variables en el uevo archivo, una sola de las variables, lo que provoca que en el archivo combinado, los casos correspondientes a la variable no pasada aparecen como perdidos del sistema.

Se puede, también, crear una nueva variable que registre el origen de los datos en el nuevo archivo combinado, para ello sólo hay que marcar la opción correspondiente en el cuadro de diálogo Indicar el origen del caso como variable. Por defecto la nueva variable se denomina origen01, pero se puede dar otro nombre, y los valores son 0 para los casos aportados por el archivo de trabajo y 1 para los casos aportados por el archivo externo.

3.6.2 Añadir variables

Para añadir a un archivo de trabajo un archivo externo con nuevas variables es preciso que ambos archivos contengan la misma variable y que en ambos estén

ordenados los casos según un criterio ascendente. Al procedimiento se accede siguiendo la secuencia:

Datos → Fundir archivos → Añadir variables

y se accede al cuadro **Abrir archivos**. En dicho cuadro se elige el archivo externo (*Datos de empleados con salario actual*) que queremos fusionar al de trabajo (*Datos de empleados con salario inicial*) y, una vez abierto, se muestra el cuadro de diálogo de la Figura 3.10.

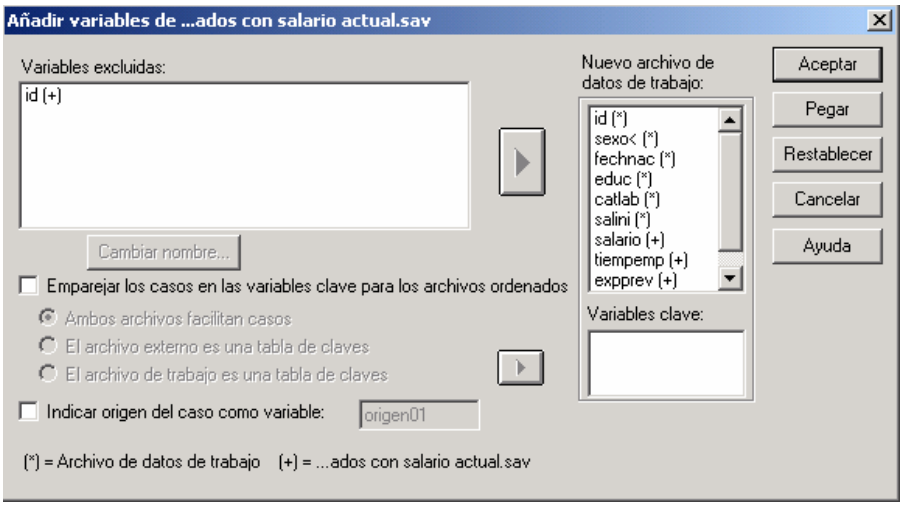


Figura 3.10. Cuadro de diálogo *Añadir variables de...*

Dado que en el archivo externo está también la variable id, el programa la excluye y la señala con el signo +, indicando que la variable es aportada por dicho archivo. En la lista **Nuevo archivo de datos de trabajo**, se muestran las variables que compondrán este nuevo archivo.

En este cuadro, lo primero es marcar la variable excluida –**id(+)**- y señalar la opción **Emparejar los casos en las variables clave para los archivos ordenados**. A continuación, se pasa la variable excluida –**id(+)**- a la lista **Variables clave**. Como ambos archivos, el de trabajo y el externo, aportan casos, se deja marcada dicha opción por defecto. Cuando se pulsa aceptar, siempre se muestra un mensaje en el que se advierte que el emparejamiento no se producirá si los archivos no están ordenados de forma ascendente por la variable clave.

No siempre los dos archivos van a contener el mismo número de casos, ni siquiera los mismos casos, aunque en ambos estén ordenados de manera ascendente por la variable clave. En estas condiciones puede interesar activar la opción **Indicar origen del caso como variable**, para que en la variable que se cree se especifique qué archivo aporta el caso. Obviamente, los casos aportados por el archivo de trabajo que no estén en el externo, serán valores perdidos del sistema y viceversa. En el cuadro siguiente, se ilustra esta situación.

	id	v1
1	1	3
2	2	5
3	3	6
4	4	2
5	5	7
6	6	8
7	7	4
8	8	6
9	9	10
10	10	2

Archivo de trabajo

	id	v2
1	10	7
2	11	8
3	12	5
4	13	10
5	14	2
6	15	9

Archivo externo

	id	v1	v2	origen01
1	1	3	,	0
2	2	5	,	0
3	3	6	,	0
4	4	2	,	0
5	5	7	,	0
6	6	8	,	0
7	7	4	,	0
8	8	6	,	0
9	9	10	,	0
10	10	2	7	1
11	11	,	8	1
12	12	,	5	1
13	13	,	10	1
14	14	,	2	1
15	15	,	9	1

Nuevo archivo después de la fusión

El archivo de trabajo contiene la variable id y la variable v1 y 10 casos, el externo contiene la variable id y la variable v2, y 6 casos. En el proceso de fusión se ha activado la opción de indicar el origen del caso, y el resultado es un archivo con 4 variables, **id**, **v1**, **v2** y **origen01** y en total 15 casos, dado que tanto el archivo de trabajo como el externo tienen un caso común, el de valor 10 en la variable **id**.

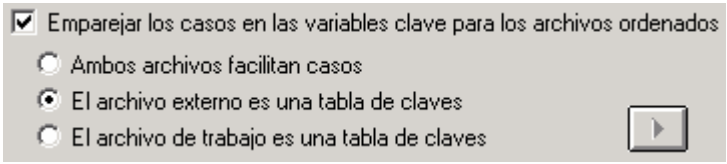
Cuando el archivo externo en vez de casos contiene una tabla de claves, el proceso es el mismo, y la única diferencia en el proceso es señalar dicha opción en el cuadro de diálogo **Añadir variables de...** El resultado del proceso de fusión es tal que cada caso del archivo externo puede ser emparejado con más de un caso del archivo de trabajo. Para ilustrar el procedimiento, emplearemos dos archivos creados *ad hoc* y que se muestran en el siguiente cuadro.

	grupo	nota
1	1	45
2	1	50
3	1	48
4	2	48
5	1	39
6	2	39
7	3	48
8	2	60
9	3	49
10	3	48
11	2	48
12	2	43

	grupo	mednota	n
1	1	45,38	4
2	2	47,61	5
3	3	48,36	3

En la parte izquierda se muestra un archivo con dos variables, **grupo** y **nota**, mientras que en la de la derecha, se muestra una tabla de claves cuyo contenido son las medias de la variable nota (**mednota**) y el número de casos por grupo (**n**). Cuando el archivo de trabajo está ordenado por la variable grupo, fusionamos

ambos archivos mediante el procedimiento **Añadir variables...** marcando en el cuadro de diálogo la opción:



el archivo resultante después de la fusión es el que se muestra a continuación:

	grupo	nota	mednota	n
1	1	45	45,38	4
2	1	50	45,38	4
3	1	48	45,38	4
4	1	39	45,38	4
5	2	48	47,61	5
6	2	39	47,61	5
7	2	60	47,61	5
8	2	48	47,61	5
9	2	43	47,61	5
10	3	48	48,36	3
11	3	49	48,36	3
12	3	48	48,36	3

en el que se observa que a cada valor de la variable de agrupamiento, **grupo**, le corresponde obviamente el mismo valor de las variables **mednota** y **n** que tenían en el archivo de claves.

3.7 Ponderar casos

Ponderar casos implica que cada registra valga más de un caso, por lo que el resultado de este procedimiento es justo el inverso del procedimiento de agregación de casos. Para ponderar casos es preciso emplear un variable de ponderación que será la que determine el valor de la frecuencia o el peso de los casos del resto de las variables con formato numérico del archivo. Su utilidad es manifiesta cuando, por ejemplo, no se dispone de los datos originales y tan sólo se tienen los datos ya agrupados y es preciso analizarlos y representarlos gráficamente, o también, en ausencia de datos originales sólo se dispone de datos de dos variables medidas conjuntamente en su forma de una distribución conjunta.

Ilustremos el proceso en primer lugar para una variable de la cual sólo se dispone de una tabla con la distribución de frecuencias que se muestra en la Tabla 3.1. En ella se muestra el número de palabras diferentes que emiten bebés de 10 meses y la frecuencia de niños que emiten ese número de palabras en la muestra de 423 bebés seleccionada. Para poder analizar estos datos, se introducen en el **Editor de datos** de SPSS de la manera habitual, como se ve en la parte derecha de la Tabla 3.1, y después se pondera el archivo, según la variable **ncasos**. De este modo, tanto los estadísticos como las representaciones gráficas de este conjunto de datos serán igual que si hubiéramos creado un archivo con una sola variable, **Nº de palabras**, con 25 ceros, 35 unos, treinta dos, etcétera.

Tabla 3.1 Distribución de frecuencias escrita en el editor de datos para posteriormente ponderar por la variable de frecuencia ncasos

Nº palabras	ncasos
0	25
1	35
2	30
3	40
4	50
5	52
6	50
7	48
8	40
9	35
10	18

	npalabr	ncasos
1	0	25
2	1	35
3	2	30
4	3	40
5	4	50
6	5	52
7	6	50
8	7	48
9	8	40
10	9	35
11	10	18

Para ponderar el archivo se sigue la secuencia:

Datos → Ponderar casos...

y se accede al cuadro de diálogo que se muestra en la Figura 3.11. En dicho cuadro se señala la opción correspondiente, y se pasa la variable que contiene los pesos o frecuencias al cuadro **Variable de frecuencia**.

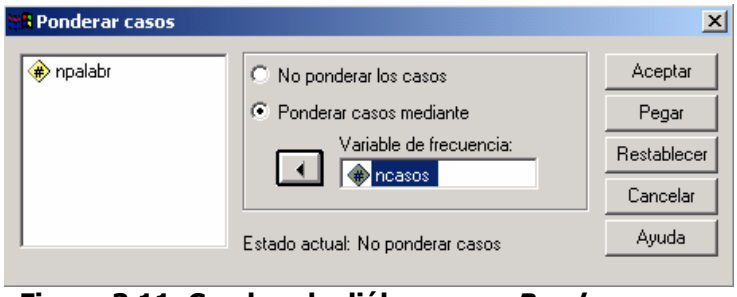


Figura 3.11. Cuadro de diálogo para *Ponderar casos*

Después de que se activa la ponderación, en el Área de estado de ponderar (esquina inferior derecha del Editor de datos[3]) aparece el aviso de que el archivo está Ponderado. Una vez ponderado, la gráfica correspondiente a la variable **npalabr**, será como se muestra en la Figura 3.12(a) y los estadísticos descriptivos los de la Figura 3.12(b).

3 Para que se vea el estado en la barra de tareas, es preciso que la resolución de la pantalla sea, al menos, de 1024 por 768 pixels.

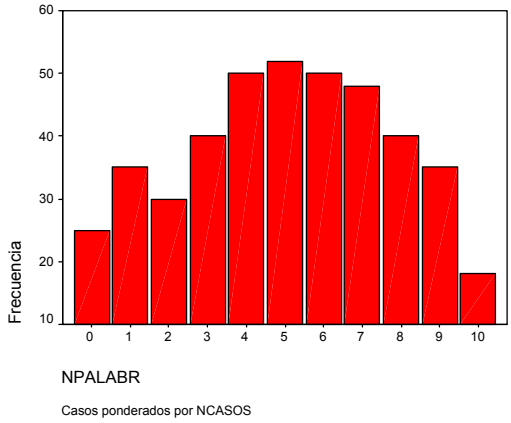


Figura 3.12 (a)

Estadísticos

NPALABR

N	Válidos	423
	Perdidos	0
Media		5,03
Desv. típ.		2,79
Asimetría		-,088
Error típ. de asimetría		,119
Curtosis		-,947
Error típ. de curtosis		,237
Percentiles	25	3,00
	50	5,00
	75	7,00

Figura 3.12 (b)

Figuras 3.12 (a) Histograma sobre un conjunto de casos ponderados; y (b) Tabla de estadísticos del conjunto de casos ponderados.

Para el caso de dos variables medidas conjuntamente, si sólo disponemos de una tabla de distribución conjunta como la que se muestra en el cuadro inferior izquierda, los datos se introducen en el editor de datos como se muestra en la parte derecha del cuadro[4].

		Y: Tipo de colegio			
		Colegio público (1)	Colegio concertado (2)	Colegio privado (3)	
X Nº de hijos	1	22	16	36	74
	2	22	26	16	64
	3	16	34	8	58
	4	12	4	0	16
		72	80	60	212

	nhijos	colegio	ncasos
1	1	Col. públ	22
2	1	Col. con	16
3	1	Col. priv	36
4	2	Col. públ	22
5	2	Col. con	26
6	2	Col. priv	16
7	3	Col. públ	16
8	3	Col. con	34
9	3	Col. priv	8
10	4	Col. públ	12
11	4	Col. con	4
12	4	Col. priv	0

Una vez ponderado el archivo por la variable **ncasos**, al invocar el procedimiento **Tabla de contingencia** (**Analizar → Estadísticos descriptivos → Tablas de contingencia...**) e incorporar la variable **nhijos** en las filas y la variable **colegio** en las columnas, el resultado es el siguiente:

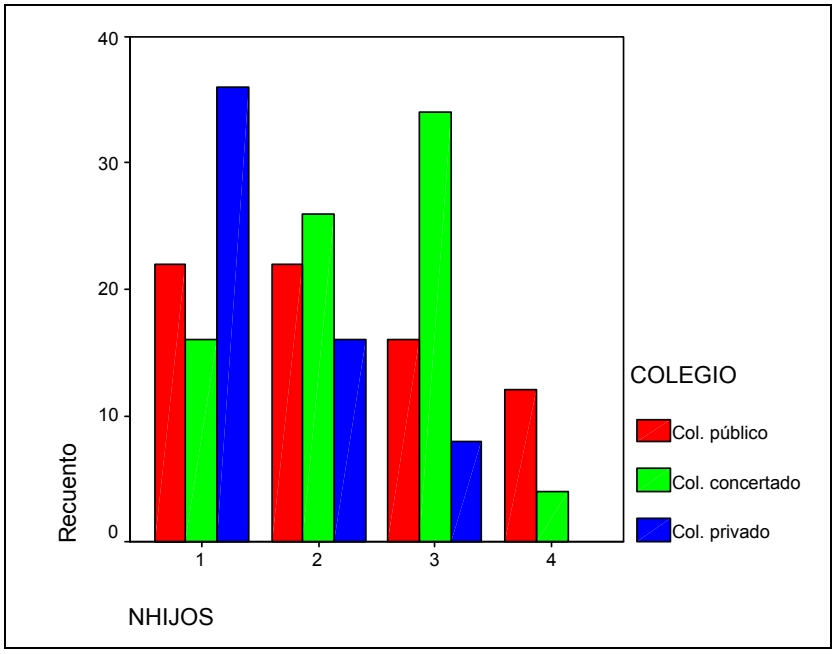
4 Aunque en la variable colegio aparecen las etiquetas (Col. público, privado, etc.) en el editor de datos se introducen los valores numéricos correspondientes a cada categoría.

Tabla de contingencia NHIJOS * COLEGIO

Recuento

		COLEGIO			
		Col. público	Col. concertado	Col. privado	Total
NHIJOS	1	22	16	36	74
	2	22	26	16	64
	3	16	34	8	58
	4	12	4		16
Total		72	80	60	212

y el gráfico de barras agrupadas que contiene dicho procedimiento sería el siguiente:



3.7 Segmentar archivo

En determinadas ocasiones puede ser útil que los resultados de nuestros análisis estén divididos de acuerdo a una o más variables categóricas. Para ello SPSS dispone del procedimiento de segmentación de archivo, al que se accede siguiendo la secuencia

Datos → Segmentar archivo...

y cuyo cuadro de diálogo es el de la Figura 3.13.

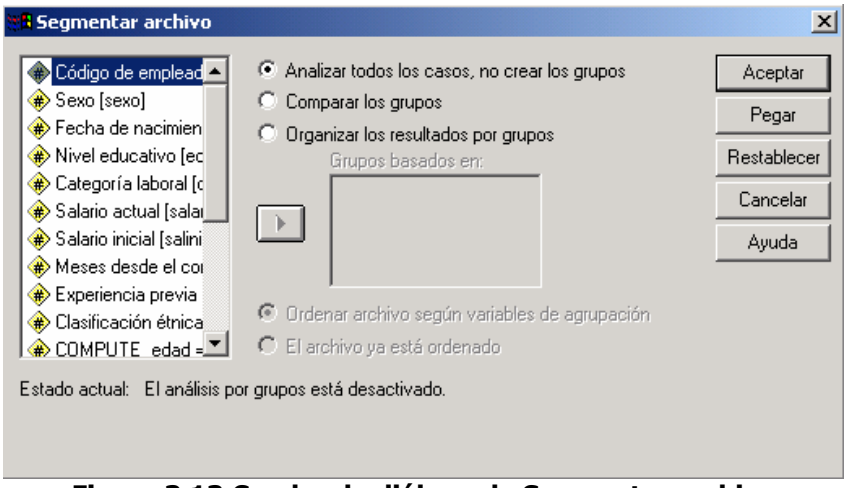


Figura 3.13 Cuadro de diálogo de Segmentar archivo

Por defecto, los datos se analizan como si formaran parte de un solo grupo, pero se dispone de dos opciones de segmentación que proporciona tablas diferentes según sea la elegida. al marcar una de las dos opciones de segmentación se activa la lista **Grupos basados en**, a la que tendremos que trasladar la/s variable/s de segmentación. Cuando un archivo está **Segmentado**, está condición se ve reflejada en la última **Área de estado**, en la parte inferior derecha del **Editor de datos**

Cuando se elige la primera opción de agrupamiento **Comparar los grupos** y pasamos a la lista **Grupos basados en** la variable **Categoría laboral**, y posteriormente se ejecuta el procedimiento descriptivos (explicado más adelante), el resultado es el que se muestra en la Tabla 3.2.

Tabla 3.2. Resultado del procedimiento Descriptivo sobre la variable Nivel educativo cuando se ha segmentado el archivo con la opción de Comparar los grupos.

Estadísticos descriptivos

Categoría laboral		N	Mínimo	Máximo	Media	Desv. típ.
Administrativo	Nivel educativo	363	8	19	12,87	2,333
	N válido (según lista)	363				
Seguridad	Nivel educativo	27	8	15	10,19	2,219
	N válido (según lista)	27				
Directivo	Nivel educativo	84	12	21	17,25	1,612
	N válido (según lista)	84				

Cuando se elige la segunda opción **Organizar los resultados por grupos**, el resultado es el que se muestra en la Tabla 3.3.

Tabla 3.3. Resultado del procedimiento *Descriptivo* sobre la variable *Nivel educativo* cuando se ha segmentado el archivo con la opción de *Organizar los resultados por grupos*.

Categoría laboral = Administrativo

Estadísticos descriptivos [a]

	N	Mínimo	Máximo	Media	Desv. típ.
Nivel educativo	363	8	19	12,87	2,333
N válido (según lista)	363				

a. Categoría laboral = Administrativo

Categoría laboral = Seguridad

Estadísticos descriptivos [a]

	N	Mínimo	Máximo	Media	Desv. típ.
Nivel educativo	27	8	15	10,19	2,219
N válido (según lista)	27				

a. Categoría laboral = Seguridad

Categoría laboral = Directivo

Estadísticos descriptivos [a]

	N	Mínimo	Máximo	Media	Desv. típ.
Nivel educativo	84	12	21	17,25	1,612
N válido (según lista)	84				

a. Categoría laboral = Directivo

4. El Visor de SPSS

4.1 Introducción

En el primer capítulo ya mencionamos que una de las novedades que presentó a partir de la versión 7, es el Visor de Resultados, interface que presenta los resultados de las operaciones que se realizan con los diferentes procedimientos. En esta ventana podemos desplazarnos con facilidad a cualquiera parte de los resultados que se han ido produciendo en las sesiones con SPSS. También se pueden modificar los resultados y crear un documento que contenga exactamente los resultados que deseemos, de manera organizada y con el formato más conveniente a nuestros propósitos.

4.2 El Visor de resultados

El Visor de resultados (Figura 4.1) se divide en dos marcos

- El marco izquierdo contiene los titulares del contenido de los resultados.
- El marco derecho contiene tablas estadísticas, gráficos y resultados de texto.
- Se pueden utilizar las barras de desplazamiento para el examen de los resultados o bien pulsar en el titular correspondiente (marco izquierdo) para ir directamente a esa tabla o gráfico.
- Se puede modificar la anchura de los marcos con sólo pulsar y arrastrar en el borde derecho del marco de titulares.

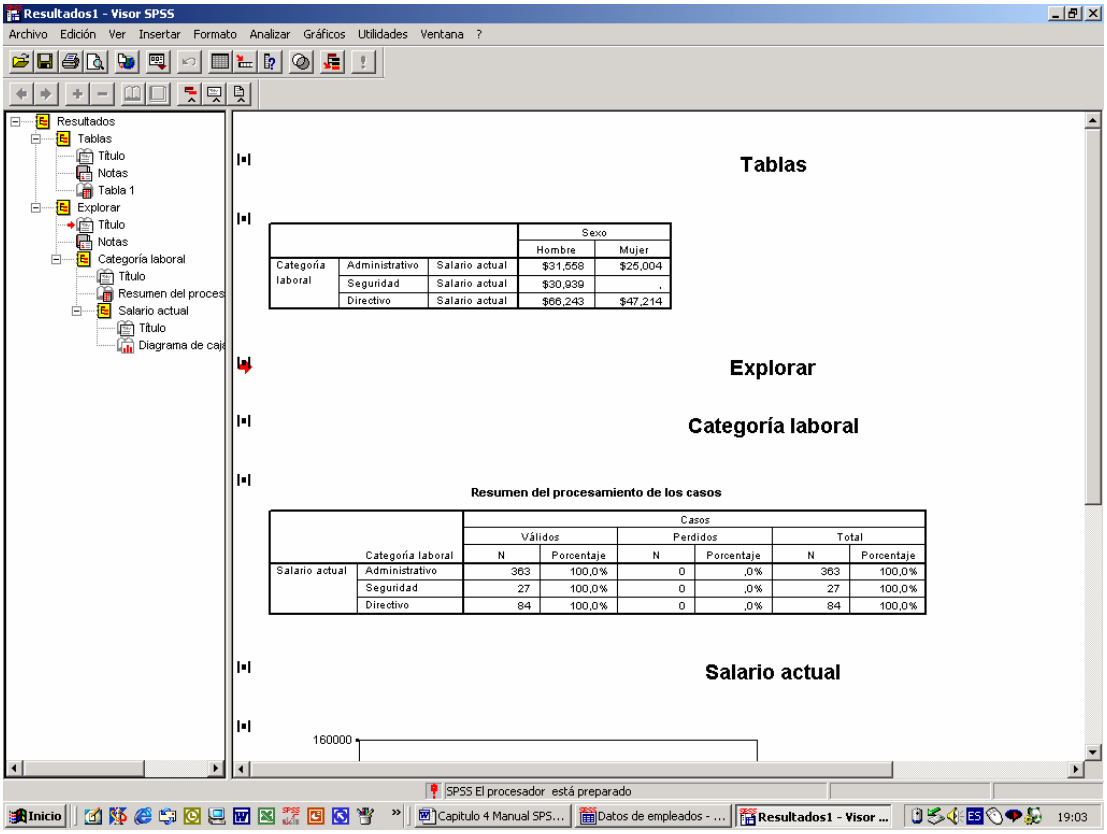


Figura 4.1 Aspecto del Visor de resultados

El contenido del Visor puede guardarse como un documento que puede ser abierto desde SPSS. El documento guardado incluye ambos marcos, el de titulares y el de resultados.

Además de las tablas estadísticas, los gráficos y los resultados de texto en el Visor se muestran otros elementos, tales como advertencias, notas y títulos. La aparición o no, en el Visor, es opcional y el usuario puede configurarlo. De manera sintética los diversas acciones que se puede realizar en el Visor son las siguientes:

- ***Almacenar el documento del Visor***. Elegir **Archivo** en su menú principal y luego **Guardar**. Por defecto, la extensión de estos documentos es SPO. También se pueden guardar los resultados en otro formato diferente mediante la opción **Exportar** en el menú Archivo.
- ***Mostrar y ocultar resultados***. De forma selectiva se pueden ocultar o mostrar las diferentes resultados que aparecen en el Visor. Para ello, se pulsa dos veces en el icono del libro del panel de titulares que corresponda a ese resultado concreto. Por defecto, por cada procedimiento requerido se despliega el resultado del mismo antecedido del título correspondiente a ese procedimiento. Si se quiere ocultar esos resultados, además del procedimiento descrito, se puede pulsar una vez en el signo menos, a la izquierda del encabezado del procedimiento, en el marco de titulares.
- ***Desplazamiento, copia y eliminación de resultados***. Para mover un resultado, se pulsa en dicho elemento en el marco de resultados y se desplaza a la posición que se desee. Para copiar, uno o varios elementos, se marcan los elementos, y en **Edición** del menú del Visor se elige **Copiar**.

Para borrar un elemento se señala el mismo y se pulsa la tecla <Suprimir>, o bien en Edición se elige **Eliminar**. Si se desea borrar un procedimiento completo, se pulsa una vez en el icono de cabecera del procedimiento y se marcarán todos los elementos, luego se pulsa <Suprimir>.

- ***Cambiar la alineación de los resultados***. Por defecto, los resultados están alineados a la izquierda. Para cambiar la alineación, se pulsa dicho elemento (en el marco de titulares o en el propio elemento) y en **Formato** del menú del Visor se elige la nueva alineación (izquierda, centro o derecha)

4.3 Tablas

La mayor parte de los resultados se presenta en formato de tablas que se pueden manipular de múltiples formas. Para ello hay que pulsar dos veces en el interior de la tabla, y ésta se edita en su propia ventana, el Editor de Tablas, cuyo aspecto es el que se muestra en la Figura 4.2, aparentemente no difiere del aspecto que muestra el Visor cuando presenta los resultados.

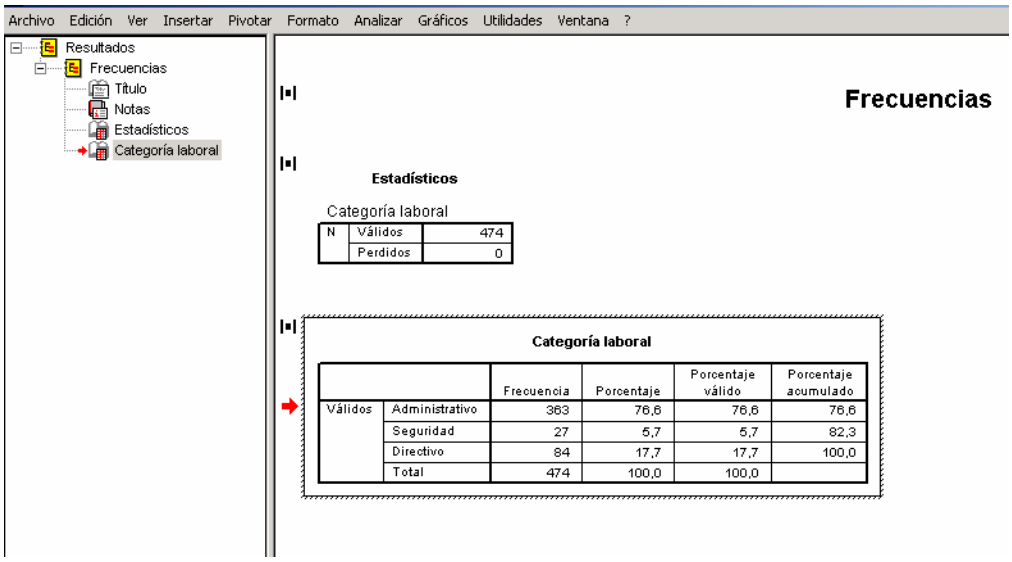


Fig. 4.2 Tabla pivote en su ventana de edición.

Una manera rápida de detectar que se está en el editor de Tablas es que el recuadro de la tabla marcada con doble clic no es un marco fino, sino con rayitas pequeñas alrededor del marco. Respecto al menú principal, aunque algunas de las opciones tienen el mismo nombre, las operaciones que se pueden hacer son diferentes. Por ejemplo, observe el lector la opción **Formato** del Visor y la opción **Formato** cuando ya se ha hecho doble clic sobre una tabla, es decir cuando se ha entrado en el **Editor de tablas**. En la opción del Visor, el Formato se refiere a la posición del objeto (tabla, gráfico, etc.) dentro de la página impresa, es decir, izquierda, centro o derecha, mientras que en esa opción una vez que se ha editado, se refiere a los diferentes cambios que se pueden efectuar en el aspecto de la tabla. En las Figuras 4.3a y 4.3b, se puede observar esta diferencia.

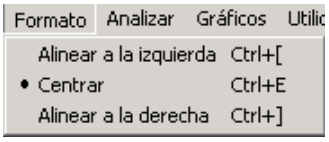


Figura 4.3(a)

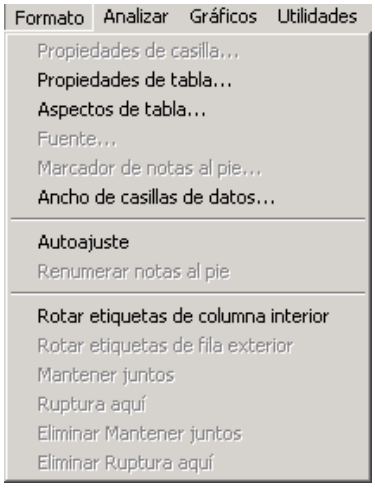


Figura 4.3 (b)

Figuras 4.3 (a) Opciones menú formato del Visor; y (b) Opciones menú formato del Editor de Tablas

Otras posibles acciones que se pueden efectuar sobre las tablas son las siguientes:

- ***Pivotaje de la tabla a través de iconos***. En el menú elegir **Pivotar** y **Paneles de pivotado**, y el aspecto del panel es el de la Figura 4.4.

Fig. 4.4 Panel de pivotado de tablas

En este panel se pueden transponer filas y columnas con sólo intercambiar los iconos correspondientes (pulsar y arrastrar).

- ***Agrupar filas o columnas e insertar etiquetas de grupo***. Activar la tabla pivote. Seleccionar las etiquetas de las filas o columnas que se quiere agrupar (pulsar y arrastrar). En el menú **Edicción** elegir **Agrupar**. Automáticamente se inserta una etiqueta de grupo cuyo texto se puede editar pulsando dos veces.
- ***Desagrupar filas o columnas y eliminar etiquetas de grupo***. Activar la tabla pivote. Seleccionar las etiquetas de las filas o columnas que se quiere desagrupar (pulsar y arrastrar). En el menú **Edicción** elegir **Desagrupar**. Automáticamente se elimina la etiqueta de grupo.
- ***Rotar etiquetas de filas o columnas***. Activar la tabla pivote. En **Formato** del menú elegir **Rotar etiquetas de columna interior** o bien **Rotar etiquetas de fila exterior**. El resultado para la columnas es el que se ve en el siguiente cuadro

		Sexo		Total
		Hombre	Mujer	
Categoría Laboral	Administrativo	157	206	363
	Seguridad	27		27
	Directivo	74	10	84
Total		258	216	474

Tablas con columnas interiores rotadas

- ***Cambio del aspecto de las tablas***. Por defecto las tablas se presentan con un formato, el cual puede ser cambiado mediante la opción **Aspectos de la tabla** en **Formato** del menú principal. Son muchas las ocpiones de presentación que se muestran en la lista de archivos de aspecto. En la ventana de la de la derecha se presenta dicho aspecto.
- ***Propiedades de la tabla***. Para establecer las propiedades de una tabla, elegir dicha opción en **Formato** del menú del editor de tablas. Se puede variar el aspecto general, las notas al pie, los formatos de las casillas y los bordes. La carpeta con las diferentes propiedades es la que se muestra en la Figura 4.5.

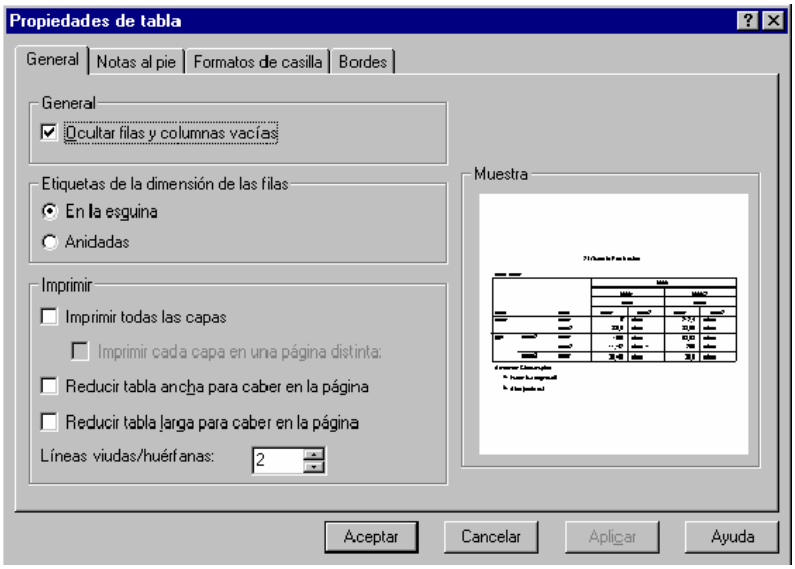


Figura 4.5 Carpeta para elegir diferentes propiedades de las tablas pivote.

- ***Fuente*** También se puede modificar la fuente para distintas areas de la tabla pivote que contienen texto. Las opciones incluyen el tipo, el estilo y el tamaño. También se puede ocultar el texto o subrayarlo. Si se especifican las propiedades de fuente en una casilla, se aplicarán en todas las capas de

la tabla que tengan la misma casilla. Para cambiar la fuente se pulsa la casilla concreta y se elige **Fuente** en el menú **Formato**.

4.4 Utilización de resultados de SPSS en otras aplicaciones

Las tablas y los gráficos de SPSS se pueden copiar y pegar en otra aplicación que corra en entorno Windows, sea un pocesador de texto o una hoja de cálculo. He aquí las operaciones a seguir.

- ***Copiar tabla o gráfico***. Se seleciona la tabla o gráfico y en **Edición** del menú del Visor se elige **Copiar**
- ***Copiar datos de una tabla pivote***. Activar la tabla. Seleccionar las etiquetas de los datos que se quieren copiar. Luego se sigue la secuencia: **Edicción**, **Seleccionar**, **Cuerpo de tabla** o **Casillas de datos** o **Casillas de datos y etiquetas**. Una vez hecha la selección elegir **Copiar** del menú de **Edicción**.
- ***Pegar los resultados en otra aplicación***. Una vez copiado/s el/los resultado/s en SPSS, elegir en el menú de **Edicción** de la aplicación de destino la opción **Pegar** o bien **Pegado especial**. En la mayor parte de las aplicaciones, **Pegar** pegará los resultados de SPPS como imagen (metaarchivo). En la Figura 4.6 puede verse la manera como se pega una tabla, en formato gráfico en un procesador de texto y en formato propio de Excel.

Tabla de frecuencia Categoría Laboral

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	Administrativo	363	76,6	76,6
	Seguridad	27	5,7	82,3
	Directivo	84	17,7	100,0
	Total	474	100,0	
Total		474	100,0	

Tabla de SPSS copiada como una imagen en un procesador de texto

	A	B	C	D	E	F
1						
2	Tabla de frecuencia Categoría Laboral					
3			Frecuencia	Porcentaje	Porcentaje acumulado	
4	Válidos	Administrativo	363	76,5822785	76,5822785	
5		Seguridad	27	5,69620253	82,278481	
6		Directivo	84	17,721519	100	
7		Total	474	100		
8	Total		474	100		
9						

Tabla pegada en Excel

Figura 4.6. Dos formas de pegado de una tabla pivote de SPSS en otra aplicación

Como puede verse, en las hojas de cálculo se pega el resultado exacto de la operación y no las cifras redondeadas que se muestran en las tablas pivote del Visor de SPSS, es decir, se copia el dato no su imagen.

Pegar especial permite seleccionar los resultados que SPSS copia en el Portapapeles en múltiples formatos. Los más frecuentes en las aplicaciones de destino son el Texto sin formato o la Imagen

4.5 Exportar resultados

En el cuadro de diálogo **Exportar resultados** se pueden guardar las tablas y los resultados de texto en formato HTLM y de texto, y los gráficos en una amplia variedad de formatos. Las posibilidades son las siguientes:

- ***Documentos de resultados***. Se puede exportar cualquier combinación de tablas pivote y gráficos. Estos últimos se exportan en el formato de exportación que esté seleccionado en ese momento. Para cada gráfico se genera un archivo distinto. Para el formato HTML los gráficos se incrustan por referencia.
- ***Documentos de resultados (sin gráficos)***. Se exportar tablas pivote y resultados de texto. Las tablas se pueden exportar como tablas HTML (3.0 o posterior), como texto separado por tabuladores o como texto separado por espacios.
- ***Gráficos***. Se pueden exportar como metaarchivo de Windows, mapa de bits de Windows, PostScript encapsulado, JPEG, TIFF, CGM, o PICT de Macintosh.

5. Sintaxis de comandos en SPSS

5.1 Introducción

Como ya señalamos en la presentación, SPSS funciona internamente por medio de un lenguaje de comandos con una sintaxis especifica, aunque la mayor parte de ellos pueden ser accesibles a través de los menús y cuadros de diálogo. Sin embargo, algunos de los comandos y opciones sólo son accesibles mediante el uso de ese lenguaje de comandos. En este capítulo explicaremos la forma de trabajar por medio del lenguaje de comandos, y la forma de generar archivos con instrucciones en este lenguaje que posibilitan la repetición posterior de los análisis de forma automática sin tener que recurrir a la selección mediante los menús.

Un archivo de sintaxis es simplemente un archivo de texto que contiene instrucciones de comandos de SPSS. Aunque se puede abrir una ventana de sintaxis y escribir comandos en ella, es mucho más sencillo indicar a SPSS que lo haga por nosotros, siempre que la operación que queremos realizar sea accesible a través de menús. En los pocos casos en que ésta no sea accesible desde los menús, no quedará más remedio que escribir las instrucciones.

Para generar unas instrucciones sin necesidad de escribirlas hay tres métodos alternativos:

- Pegar la sintaxis de comandos desde los cuadros de diálogo
- Copiar la sintaxis desde el registro de resultados
- Copiar la sintaxis desde el archivo diario

En la Ayuda en pantalla de un procedimiento determinado de SPSS, pulsando el botón de Sintaxis se puede saber qué opciones del lenguaje de comandos están disponibles para ese procedimiento y acceder al diagrama de sintaxis de ese comando concreto.

5.2 Creación de instrucciones desde los cuadros de diálogo

Es el método más sencillo de generar un archivo de sintaxis de comandos y consiste en pulsar el botón **Pegar** del cuadro de diálogo una vez se hayan realizado las selecciones específicas para ese procedimiento concreto. Por ejemplo, en la Figura 5.1 se muestra el texto de las instrucciones correspondientes al procedimiento de Frecuencias con la especificación de generar un diagrama de barras (BARCHART) que refleje las frecuencias (FREQ) de las variables **catlab** y **minoria**, que se han insertado-pegado en la ventana de sintaxis cuando, después de hecha la selección de opciones, se ha pulsado el botón **Pegar**

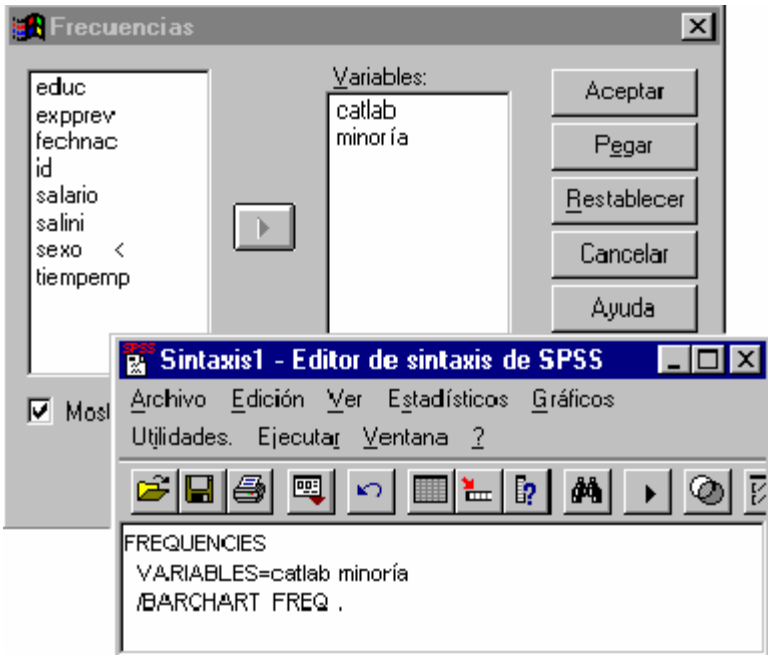


Figura 5.1 Sentencias del procedimiento Frecuencias pegado en la ventana de sintaxis

5.3 Copiar desde el registro de resultados

Por defecto, las instrucciones específicas para ejecutar un procedimiento no se muestran en el Visor, pero puede modificarse esta opción seleccionando *Mostrar los comandos en el registro* en la pestaña del Visor del cuadro de diálogo **Opciones de SPSS**, del menú **Edicción**, del que hablaremos más adelante.

Cuando está activada esta opción, en la ventana del Visor de resultados se muestra el texto de la sintaxis para ese procedimiento, como puede verse en la Figura 5.2.

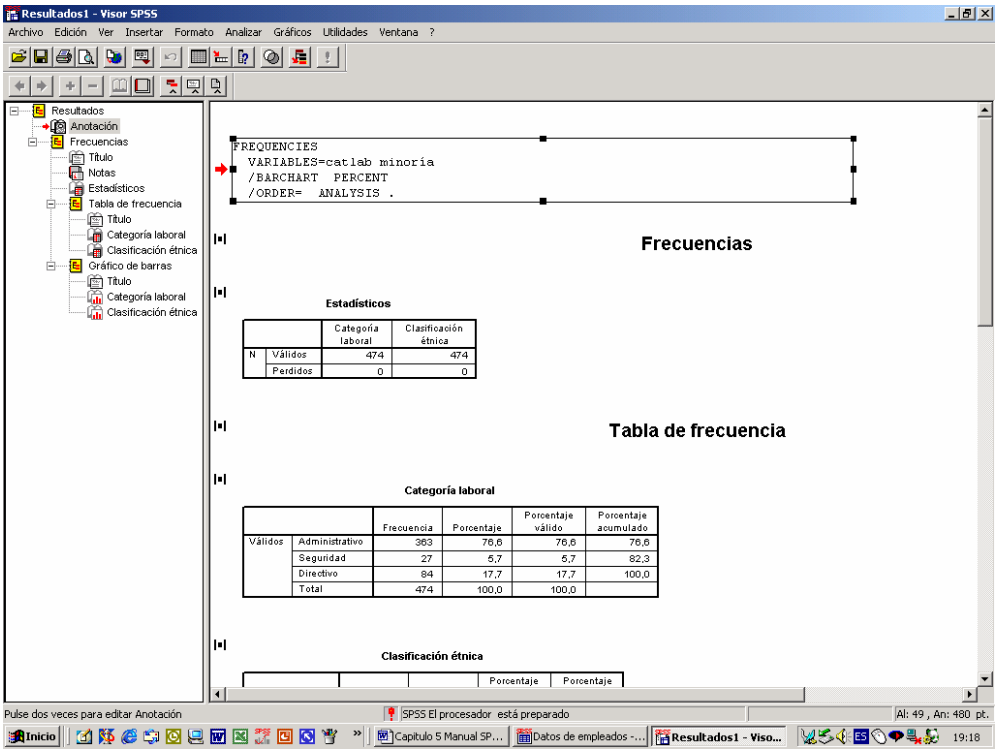


Figura 5.2 Sentencia del procedimiento Frecuencias insertadas en el Visor.

Para copiar la sintaxis hay que marcarla, en el panel de titulares, pulsando el icono del libro denominado *Anotación*, ya que la sintaxis es simplemente texto. Una vez marcado (se recuadra el contenido y se señala con una flecha) en **Edicción** del menú del Visor se elige **Copiar**. Posteriormente, se pega en una ventana de sintaxis, mediante la secuencia **Edicción** - **Pegar** del menú de esa ventana.

Se pueden copiar y pegar tantas secuencias de comandos de procedimientos como se desee. Para ello, se señala en los titulares *Anotación* correspondientes y se copian del mismo modo que si fuera uno. Luego se pega por el procedimiento ya señalado.

5.4 Copiar desde el archivo diario

Todos las operaciones que realizamos en una sesión de trabajo con SPSS son guardadas en un archivo de trabajo diario denominado SPSS.JNL. Por defecto las instrucciones de cada sesión con SPSS sobreescriben las de la sesión precedente, pero es posible modificar esto para que las operaciones de las sesiones se añadan unas a continuación de otras.

Por defecto, este archivo se almacena en C:\WINDOWS\TEMP\, pero se puede especificar otra ruta. El único incoveniente es que en este archivo diario se graba todo: las instrucciones, los mensajes de error y las advertencias que emite SPSS cuando hemos cometido alguna infracción de las normas de funcionamiento de SPSS (en la Figura 7.3 se muestra el texto de error al haber intentado obtener una distribución de frecuencias de una variable, **pepe**, que en realidad no existe en el archivo de trabajo). Por tanto, para usar la sintaxis, habrá de depurar el archivo de

esos mensajes y guardar sólo las instrucciones. Para guardarlo como archivo de sintaxis, conviene especificarle la extensión propia de estos archivos (.SPS).

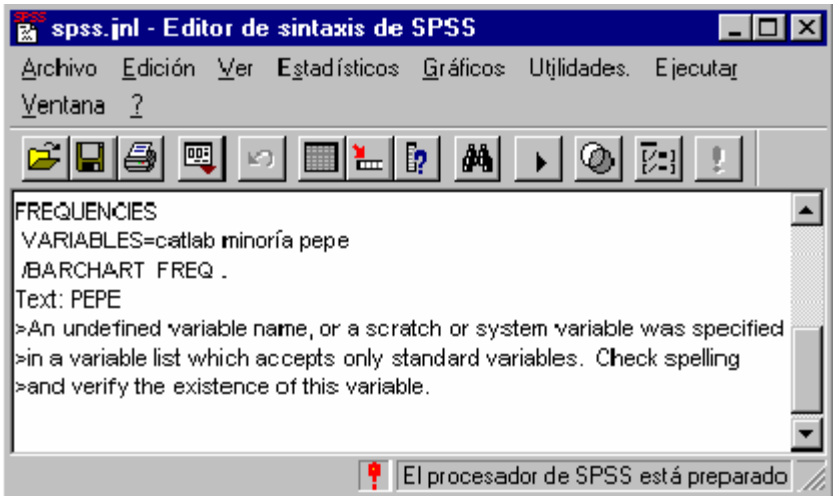


Figura 5.3 Mensaje de error insertado en el archivo de trabajo diario de SPSS

5.5 Ejecución de la sintaxis de comandos

Para ejecutar un procedimiento escrito en la ventana de sintaxis, hay que marcar todo el procedimiento y o bien pulsar el botón Ejecutar  , o bien seleccionar de entre la opción del menú Ejecutar la que consideremos más adecuada. Las posibles acciones son las siguientes:

- **Todo**. Ejecuta todos los comandos de la ventana de sintaxis
- **Selección**. Ejecuta los comandos seleccionados, incluidos los comandos resaltados parcialmente.
- **Actual**. Ejecuta el comando donde se encuentra el curso
- **Hasta el final**. Ejecuta todos los comandos incluidos desde la posición actual del cursor hasta el final del archivo de sintaxis de comandos.

5.6 Reglas básicas de la sintaxis de comandos

Las siguientes reglas han de tenerse en cuenta a la hora de escribir las sintaxis de los comandos:

- Cada comando debe empezar en una línea nueva y terminar con un punto (.).
- La mayoría de los subcomandos están separados por barras inclinadas (/). La que precede al primer subcomando de un comando generalmente es opcional.
- Los nombres de las variables deben escribirse completos.
- El texto entre comillas o apóstrofos debe contenerse en una sola línea.

- Las líneas de sintaxis no puede exceder los 80 caracteres.
- Los decimales se indican con el punto (no la coma) independientemente de la configuración regional de Windows.

La sintaxis de comandos de SPSS no distingue entre mayúsculas o minúsculas y se puede usar las abreviaturas de tres letras para designar los comandos. Daría igual escribir

```
FREQUENCIES
 VARIABLES = CATLAB MINORIA
 /BARCHART.
```

que escribir

```
fre var=catlab minoria/bar.
```

Se pueden usar tantas líneas como se desee para especificar un sólo comando. Se pueden añadir saltos de línea en casi cualquier punto donde se permite un espacio en blanco (alrededor de las barras inclinadas, los paréntesis, los operadores aritméticos o entre los nombres de las variables).

Para una mejor comprensión posterior de lo que la secuencia de instrucciones lleva a cabo, es conveniente introducir, intercalados, comentarios explicativos. Todos los comentarios deben ir precedidos de un asterisco o de la palabra COMMENT, y debe terminar con un punto al final de la última línea, por lo cual, en los comentarios no se deben introducir puntos intermedios, porque SPSS lo interpretaría como final de comentario, y las frases posteriores provocarían errores. Por ello, las pausas entre frases se señalarán con punto y coma, dos puntos, coma, pero nunca con un punto, que se reserva para el final de la instrucción.

6. Opciones de SPSS y personalización de menús

6.1 Introducción

Como ya hemos señalado, SPSS tiene una serie de opciones por defecto que el usuario puede cambiar siempre que lo desee. Además, también por defecto, en la barra de herramientas situada debajo del menú principal de cada ventana (Editor de datos, Visor, etc.) aparecen una serie de iconos por defecto que permite el acceso rápido a determinadas funciones. El usuario puede ampliarla a voluntad o suprimir algunos o todos los que se muestran. También se pueden generar nuevas barras de menús.

6.2 Opciones de SPSS

Se puede acceder a la carpeta de opciones de SPSS desde el menú **Edicción** del Editor de datos o desde el mismo menú en la ventana del Visor, y eligiendo **Opciones**. En las carpetas que se muestran en la Figura 6.1 están contenidas las diferentes opciones, y a cada una se accede pulsando la correspondiente pestaña. En este manual sólo comentaremos las opciones de la carpeta general, la del Visor y la relativa a las Tablas pivote, y sugerimos al lector que explore las posibilidades del resto de las pestañas de opciones.

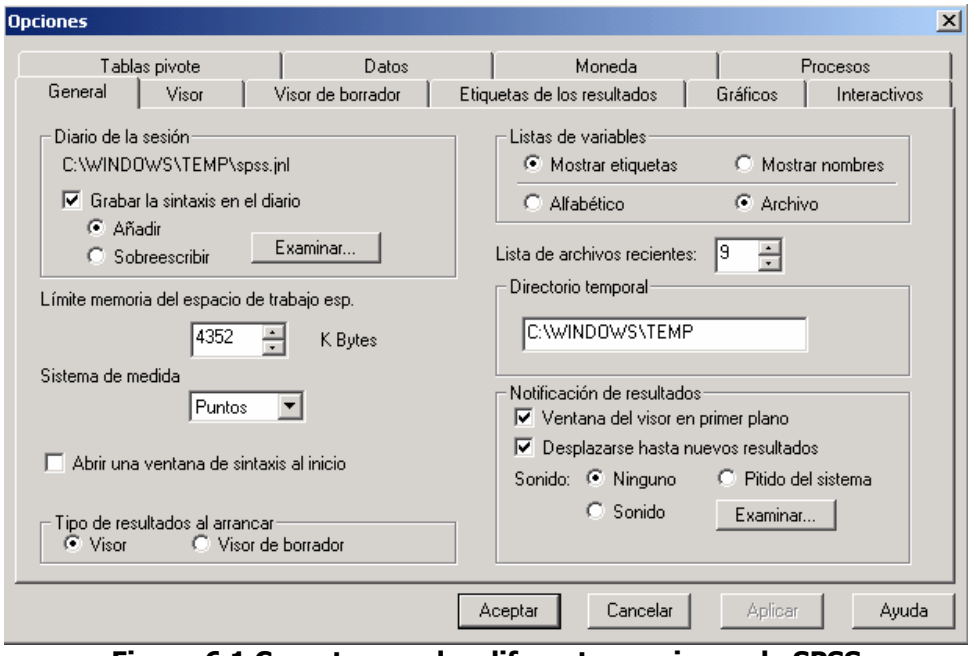


Figura 6.1 Carpetas con las diferentes opciones de SPSS

Las opciones de la carpeta denominada **General** son las siguientes:

- **Diario de la sesión**. SPSS crea y mantiene automáticamente un archivo diario de todos los comandos que se ejecutan en una sesión de SPSS. Esto incluye comandos introducidos y ejecutados en ventanas de sintaxis y

comandos generados por elecciones de cuadros de diálogo. Puede editar el archivo diario y volver a utilizar los comandos de nuevo en otras sesiones de SPSS. Puede activar o desactivar el registro de sesión, añadir o sobrescribir el archivo diario, y seleccionar el nombre y la ubicación del mismo. Puede copiar la sintaxis de comandos del archivo diario y guardarla en un archivo de sintaxis para su uso con la unidad de producción automatizada de SPSS.

- **Límite de memoria del espacio de trabajo especial**. La memoria de trabajo se puede asignar según se necesite durante la ejecución de la mayoría de los comandos. Sin embargo, existen ciertos procedimientos que requieren todo el espacio de trabajo disponible al comienzo de la ejecución. Entre los procedimientos que podrían requerir todo el espacio de trabajo disponible durante su ejecución se encuentran Frecuencias, Tablas de contingencia, Medias y Pruebas no paramétricas. Si recibe un mensaje que indica que debería cambiar la asignación del espacio de trabajo, aumente el límite de memoria especial del espacio de trabajo. Para decidir sobre un nuevo valor, utilice la información que se muestra en la ventana de resultados antes del mensaje de falta de memoria. Una vez que haya terminado con el procedimiento, probablemente deberá reducir el límite a su cantidad original (por defecto 512K), ya que un aumento de la asignación del espacio de trabajo podría reducir el rendimiento bajo ciertas circunstancias.
- **Abrir ventana de sintaxis al inicio**. Las ventanas de sintaxis son ventanas de archivos de texto utilizadas para introducir, editar y ejecutar comandos de SPSS. Si trabaja frecuentemente con la sintaxis de comandos, seleccione esta opción para abrir automáticamente una ventana de sintaxis al principio de cada sesión de SPSS. Esto es útil primordialmente para usuarios de SPSS con experiencia que prefieran trabajar con la sintaxis de comandos en vez de con los cuadros de diálogo.
- **Sistema de medida**. Sistema de medida utilizado (puntos, pulgadas o centímetros) para especificar atributos tales como los márgenes de casillas de las tablas pivote, los anchos de casilla y el espacio entre las tablas para la impresión.
- **Mostrar orden de listas de variables**. Las variables se pueden mostrar en orden alfabético o por orden según el archivo, que es el orden en el que figuran realmente en el archivo de datos (y en el que se muestran en el **Editor de datos**). Los cambios en el orden de visualización tendrán efecto la siguiente vez que se abra un archivo de datos. El orden de visualización afecta sólo a las listas de variables de origen. Las listas de variables de destino siempre reflejan el orden en el que las variables han sido seleccionadas.
- **Lista Archivos utilizados recientemente**. Controla el número de archivos utilizados recientemente que aparecen en el menú Archivo.
- **Notificación de resultados**. Controla la manera en la que SPSS notifica que se ha terminado de ejecutar un procedimiento y que los resultados están disponibles en el Visor.

Las opciones de visualización de los resultados del **Visor** sólo afectan a los nuevos resultados que se producen después de cambiar las selecciones. Los resultados mostrados previamente en el Visor de resultados no se verán afectados por los cambios de estas selecciones. Su aspecto es el de la Figura 8.2.

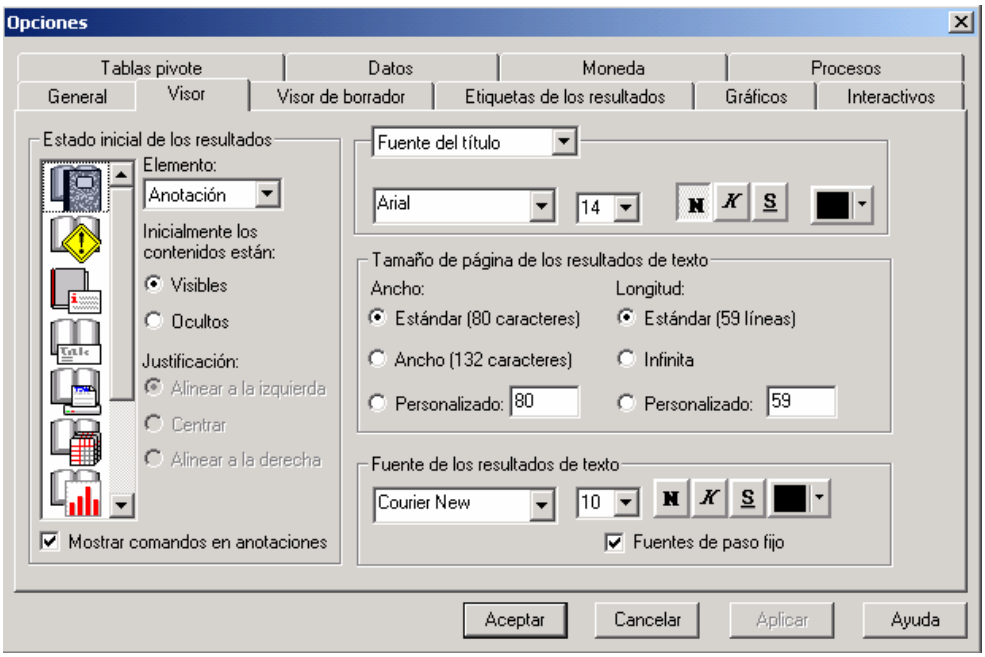


Figura 6.2 Opciones del Visor de resultados

y las opciones disponibles son las siguientes:

- **Estado inicial de los resultados**. Controla los elementos que se muestran y se ocultan automáticamente cada vez que se ejecuta un procedimiento además de su alineación inicial. Puede controlar la visualización de los siguientes elementos: anotaciones, registro, advertencias, notas, títulos, tablas pivote, gráficos y resultados de texto (los resultados no se muestran en formato de tabla pivote). También se puede activar o desactivar la visualización de los comandos de SPSS en el registro.

  Nota: Todos los elementos de los resultados se muestran alineados a la izquierda en el Visor de resultados. Únicamente se verá afectada por las selecciones de justificación la alineación de los resultados impresos. Los elementos centrados y alineados a la derecha se identifican por un pequeño símbolo en la parte superior y a la izquierda del elemento.

- **Fuente del título**. Controla el estilo, el tamaño y el color de la fuente de los nuevos títulos de resultados.

- **Tamaño de página de los resultados de texto**. En los resultados de texto, controla el ancho de página (expresado en número de caracteres) y el largo de página (expresado en número de líneas). En algunos procedimientos, algunos estadísticos se muestran sólo en formato ancho.

- **Fuentes de los resultados de texto**. Fuente utilizada para los resultados de texto. Los resultados de texto que muestra SPSS están diseñados para su utilización con fuentes de paso fijo. Si selecciona una fuente que no sea de paso fijo, los resultados tabulares no se alinearán adecuadamente.

Opciones de SPSS

Las opciones de las **Tablas pivote** para los resultados son las que se muestran en la Figura 6.3. Aquí se puede elegir un aspecto de tabla entre los varios que hay en la lista **Aspecto de Tabla**. Una vez elegido un aspecto específico, todas las tablas que se generen a partir de ese momento tendrán ese aspecto

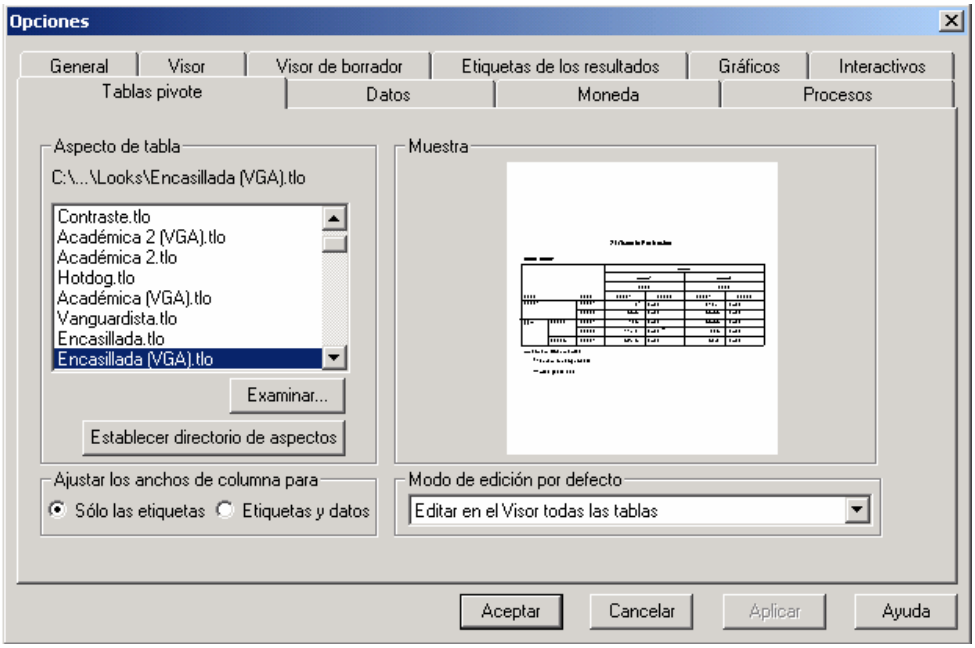


Figura 6.3 Opciones sobre el aspecto de las tablas pivote

6.3 Personalización de barras de herramientas

SPSS, por defecto, presenta una serie de barras de herramientas, compuestas por iconos, en cada una de las ventanas que conforman el conjunto del programa (Editor de datos, Visor, Gráficos, Tablas pivote, etc.). El conjunto de iconos que componen cada barra de herramientas es restringido, pero lo podemos ampliar a voluntad, e incluso crear nuevas barras de herramientas que incorporen los iconos de operaciones que deseemos y que se activen en la ventana que queramos. El procedimiento es muy sencillo, y aquí lo vamos a explicar de forma somera.

Se accede a la configuración de las barras de herramientas a través de la opción **Ver** tanto del menú del Editor de Datos como del Visor, si se accede desde el primero se muestra la configuración de la barra de herramientas del Editor, tal como se muestra en la Figura 6.4.

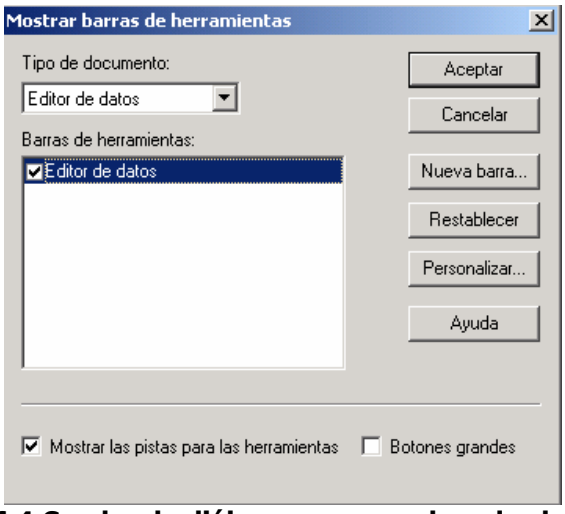


Figura 6.4 Cuadro de diálogo para acceder a las barras de herramientas de las diferentes ventanas de SPSS.

Pulsando en el despleglable **Tipo de documento** se acceden a todas las barras de todas las ventanas que hay configuradas. Para modificar el contenido de iconos de esta barra del **Editor de Datos**, pulsamos el botón **Personalizar** y accedemos a un cuadro como el que se muestra en la Figura 6.5.

Figura 6.5 Cuadro para personalizar la barra de herramientas del Editor de datos

En la ventana de la izquierda se muestran las diferentes categorías (Archivo, Edición, Ver, ...), y en la lista de la derecha los elementos de cada categoría, con sus iconos respectivos, entre los que se puede elegir para que aparezcan en esa barra de herramientas. En la parte de abajo se muestra cuáles son los iconos que actualmente configuran la barra.

El procedimiento para incorporar elementos de una categoría determinada es muy sencillo: primero se pulsa en la categoría deseada, y en la ventana de elementos aparecen los que son propios de esa categoría. En esta ventana pulsamos el icono de la función que deseemos, y lo arrastramos (con el botón izquierdo del ratón pulsado) hacia los iconos que componen la actual configuración

de esa barra y se inserta en lugar que deseemos. Es conveniente insertar iconos de separación para distinguir entre los elementos de cada categoría. Si queremos podemos cambiar el aspecto de algunos de los iconos (no todos son modificables) que estén actualmente seleccionados para componer la barra de herramientas. Para ello se pulsa dicho icono en la **Personalización de la Barra** y luego se pulsa el botón **Editar herramienta**, entrando así, en el editor de mapas de bits que incorpora SPSS (con un aspecto parecido al del Paintbrush que incorpora como accesorio Windows), pudiendo entonces cambiar su aspecto.

Por ejemplo, si pulsamos en el icono de Imprimir (representado por una impresora), y a continuación pulsamos en el botón de Herramienta de edición el aspecto del mapa de bits de dicho icono es el que muestra en la Figura 6.6.

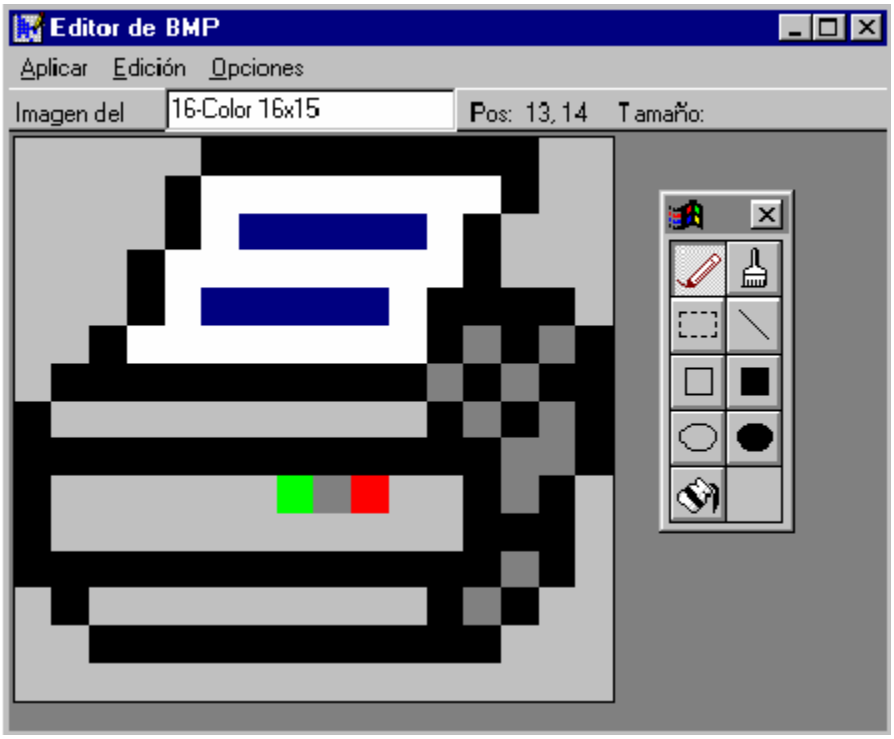


Figura 6.6 Icono de impresora en el editor de mapas de bits.

Obviamente, igual que pueden incorporarse iconos a la barra pueden eliminarse los que estén seleccionados. Basta para ello con pulsar el icono en la barra y arrastrarlo a la ventana de elementos.

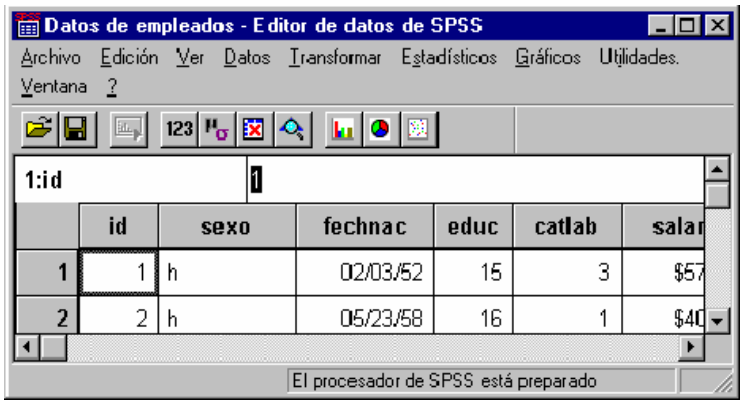


Figura 6.7 Editor de datos con iconos de funciones incorporados por el usuario

En la Figura 6.7 puede verse el aspecto de la barra de herramientas que hemos configurado, por el procedimiento señalado, con dos icono de *edición* (abrir archivo y guardar), uno de la categoría *ver* (ir a archivo gráfico), cuatro de la categoría *estadísticos* (frecuencias, descriptivos, tablas de contingencia y explorar) y tres de *gráficos* (diagrama de barras, diagrama de sectores y diagrama de dispersión). Si se quiere volver a la configuración original por defecto, sólo hay que pulsar el botón **Restablecer barra**, de modo que sugiero al lector que incorpore los iconos cuya función más va a emplear, ya que siempre puede volver al sitio de partida, mediante el botón **Restablecer barra**. En el botón **Propiedades** del cuadro de Personalización de barra de herramientas, se especifica en cuál de las ventanas deseamos que aparezca esa barra en cuestión.

Siguiendo el mismo procedimiento se pueden crear nuevas barras que contengan sólo los icono de las funciones que deseemos y que aparezca en una o varias de las diversas ventanas de SPSS. Le propongo al lector que trate de generar una nueva barra, y llámela como guste, con los iconos de los procedimientos estadísticos

- Tablas de contingencia
- Anova de un factor
- Regresión Lineal,

y los siguientes iconos de gráficos:

- Diagrama de Líneas
- Diagrama de áreas
- Histograma,

y que se muestre, junto con la que se muestra por defecto, en las ventanas del **Editor de datos** y del **Visor**.

SEGUNDA PARTE

ANÁLISIS ESTADÍSTICO

7. Análisis descriptivo

7.1 Introducción

Hay dos procedimientos básicos que permiten describir las tres propiedades de las distribuciones: la tendencia central, la dispersión y la forma –el sesgo y el apuntamiento. Además de resumir mediante índices estas propiedades, también se puede elaborar un conjunto de diagramas que permite al analista visualizar la distribución. Estos dos procedimientos que vamos a tratar en este capítulo son **Frecuencias** y **Descriptivos**.

7.2 Frecuencias

Este procedimiento (FREQUENCIES en la sintaxis del lenguaje de comandos) permite obtener distribuciones de frecuencia, estadísticos descriptivos, y gráficos de diverso tipo.

Se accede mediante

Analizar → Estadísticos descriptivos → Frecuencias...

en cualquiera de las ventanas en que aparece este menú, y se muestra el cuadro de diálogo de la Figura 7.1.

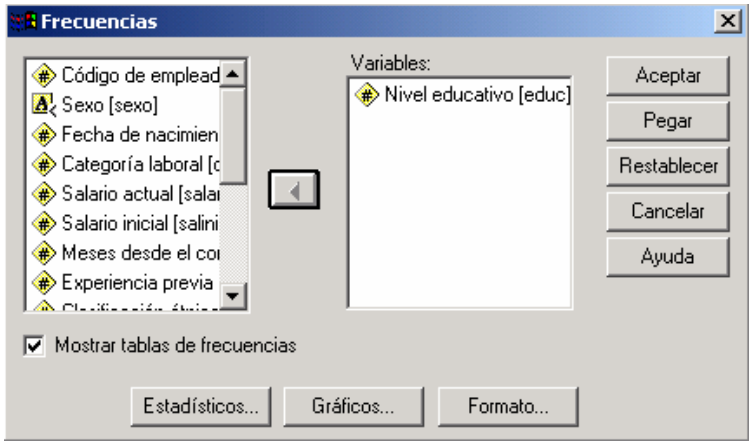


Figura 7.1 Cuadro de diálogo de *Frecuencias*

Se puede elegir los **Estadísticos** y los **Gráficos** pulsando en el botón correspondiente, así como el **Formato** de visualización de las tablas de frecuencia. Los cuadros de diálogo son los que se muestran en las Figura 7.2 (a), 7.2 (b) y 7.2 (c)

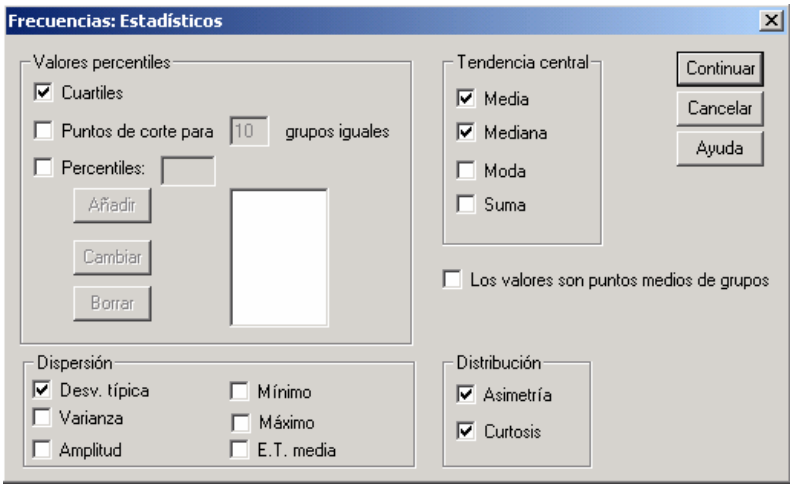


Figura 7.2 (a)

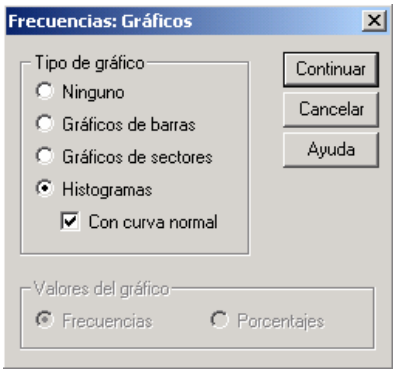


Figura 7.2 (b)

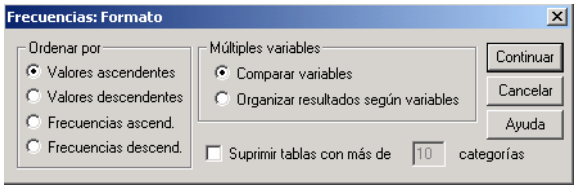


Figura 7.2 (c)

Figura 7.2 (a) Cuadros de diálogo de las opciones de Estadísticos, (b) Gráficos y (c) Formato de *Frecuencias*

Para el Histograma de frecuencias se pueden elegir que aparezca sobreimpresa la curva normal y poder, así, juzgar mejor la normalidad de los datos. En la Figura 7.3 se ve el Histograma de la variable **educ** con la cuva normal y se aprecia una clara asimetría de los datos.

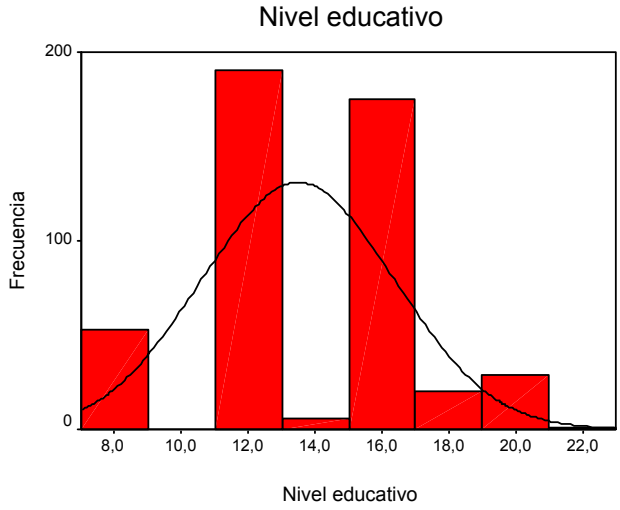


Figura 7.3 Histograma de la variable educ (nivel educativo)

Además del histograma, también se ha pedido una serie de índices estadísticos y la distribución de frecuencias de los valores de la variables. Los resultados son los que se muestran en la Tabla 7.1.

Tabla 7.1. Estadísticos y distribución de frecuencias de la variable nivel educativo

Estadísticos

Nivel educativo

N	Válidos	474
	Perdidos	0
Media		13,49
Mediana		12,00
Desv. típ.		2,88
Asimetría		-,114
Error típ. de asimetría		,112
Curtosis		-,265
Error típ. de curtosis		,224
Percentiles	25	12,00
	50	12,00
	75	15,00

Nivel educativo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	8	53	11,2	11,2	11,2
	12	190	40,1	40,1	51,3
	14	6	1,3	1,3	52,5
	15	116	24,5	24,5	77,0
	16	59	12,4	12,4	89,5
	17	11	2,3	2,3	91,8
	18	9	1,9	1,9	93,7
	19	27	5,7	5,7	99,4
	20	2	,4	,4	99,8
	21	1	,2	,2	100,0
	Total	474	100,0	100,0	

7.2.1 Estadísticos

- **Valores percentiles**. Hay varias opciones para los índices de posición. Se pueden pedir los cuartiles, especificar los percentiles que se desee y determinar *k*-1 puntos de corte para partir la distribución en *k* grupos del mismo tamaño
- **Tendencia central**. En total hay 4 índices de centralidad: Media, Mediana, Moda y Suma (que es la suma de todos los valores y por tanto el numerador del estadístico Media). Obviamente la elección del índice estará en función del tipo de variable que estemos describiendo.
- **Dispersión**. Los índices de dispersión son la varianza muestral (suma de las desviaciones cuadráticas de cada valor respecto de la media dividido por el *n*-1). La desviación típica, que es la raíz cuadrada de la varianza muestral, los valores mínimo y máximo, la amplitud y el error típico de la media, que es la desviación típica de la distribución muestral de la media, y se obtiene dividiendo la desviación típica de la muestra por la raíz cuadrada del número de casos.

- **Distribución**. Son dos los estadísticos de forma. Asimetría, que índica el sesgo de la distribución; un valor positivo indica que los valores más extremos se encuentran por encima de la media, y viceversa. También se muestra el error típico del índice de asimetría (el error típico de la distribución muestral de este estadístico), que permite tipifica el valor del índice e interpretarlo como un valor z con una distribución N(0,1). Índices tipificados mayores de 1,96 informa de una distribución asimétrica. Respecto de la Curtosis, es el índice que expresa el grado en que una distribución acumula casos en sus colas comparado con los casos que se acumulan en las colas de una distribución normal con la misma varianza. Un valor positivo indica que las colas acumulan más casos que en la normal (distribución puntiaguda), e índices próximos a cero indican una semejanza con la normal. También se muestra el error típico de la distribución muestral de la curtosis que permite interpretarlo como un valor z con distribución N(0,1).
- **Los valores son puntos medios de grupos**. Si la variable objeto de estudio está agrupada en intervalos, esta opción permite calcular los índices de posición, mediana y percentiles interpolando valores, es decir, considerando que los casos se distribuyen de forma homogénea dentro del intervalo.

7.2.2 Gráficos

El procedimiento Frecuencias ofrece algunos gráficos, tanto para variables cualitativas como para variables cuantitativas, discretas o continuas. Al pulsar el botón **Gráficos** del cuadro de diálogo **Frecuencias**, se muestra el cuadro de la Figura 7.4.

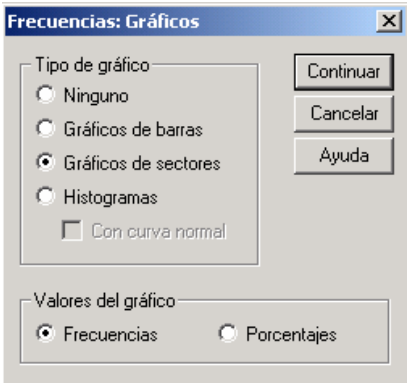


Figura 7.4 Cuadro de Gráficos de Frecuencias

De los tres tipos de gráficos, ya se ha mostrado el Histograma; en esta ocasión se ha solicitado el gráfico de sectores, expresados los valores en porcentajes. Para la variable **Categoría laboral** dicho gráfico es el que se muestra en la Figura 7.5.

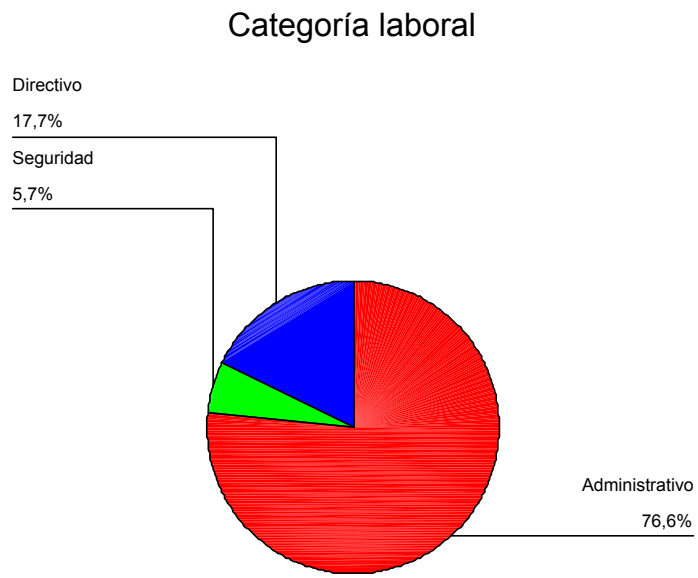


Figura 7.5 Gráfico de sectores de Categoría laboral expresado en porcentajes

Tanto el gráfico de sectores como el de barras son intercambiable, y por consiguiente, se habría podido representar la variable mediante este último. La Figura 7.6 muestra dicho gráfico.

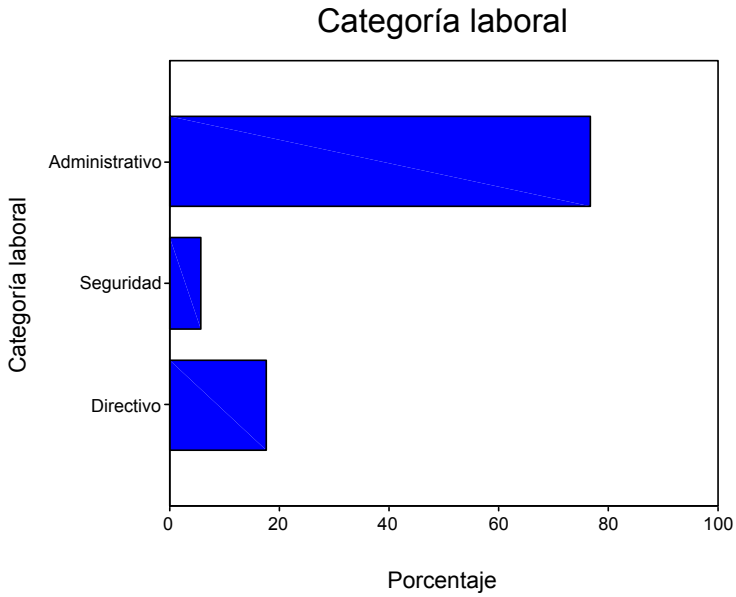


Figura 7.6 Gráfico de barras de Categoría laboral expresado en porcentajes

7.3 Descriptivos

Este procedimiento (DESCRIPTIVES) está diseñado para variables cuantitativas continuas, a diferencia del procedimiento Frecuencias que contiene opciones para todo tipo de variables. Como las opciones de estadísticos (a las que se accede

mediante el botón Opciones) son similares a las del procedimiento Frecuencias, sólo comentamos la posibilidad que ofrece este procedimiento de **Guardar los valores tipificados como variables**, o lo que es igual, el procedimiento tipifica la variable, es decir convierte las puntuaciones directas en típicas o puntuaciones z, que expresan el número de desviaciones típicas que cada valor se aleja de su media. La nueva variable guardada no es preciso darle nombre, sino que SPSS toma el valor de la variable de salida y le antepone la letra z.

Para acceder al procedimiento

Analizar → Estadísticos descriptivos → Descriptivos...

y se muestra el cuadro de diálogo que de la Figura 7.7 (a), y al pulsar el botón **Opciones** se muestra el cuadro de la Figura 7.7 (b), y en él se especifican los estadísticos que contiene el procedimiento.

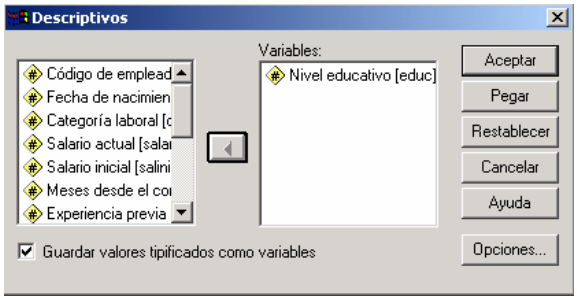


Figura 7.7 (a)

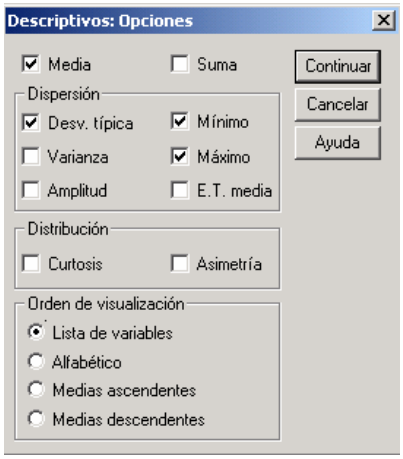


Figura 7.7 (b)

Figuras 7.7 (a) Cuadro de diálogo de Descriptivos y (b) Opciones del procedimiento

7.4 Puntuaciones típicas y curva normal

Muchas de la variables que se estudian en la ciencia en general se distribuyen normalmente. Además, según demuestra el ***Teorema central del límite***, si una variable es el resultado de la suma de un cierto número de variables independientes entre, cada una con un efecto parcial, siempre que la desviación típica de esos efectos sea finita, la distribución de esa variable se asemejará más y más a la curva normal cuanto mayor número de datos registremos, con independencia de la distribución de los efectos parciales.

La curva normal es algo así como "la madre de casi todas las distribuciones", pues de ella parten la mayoría: ji cuadrado, t de Student, F de Snedecor, etc. y a ella convergen cuando el tamaño de la muestra es elevado, de ahí su importancia en el ámbito de la ciencia en general y de la Psicología en particular. Su forma y las proporciones asociadas a determinadas puntuaciones se pueden ver en el figura 7.8, y sus características más notables son:

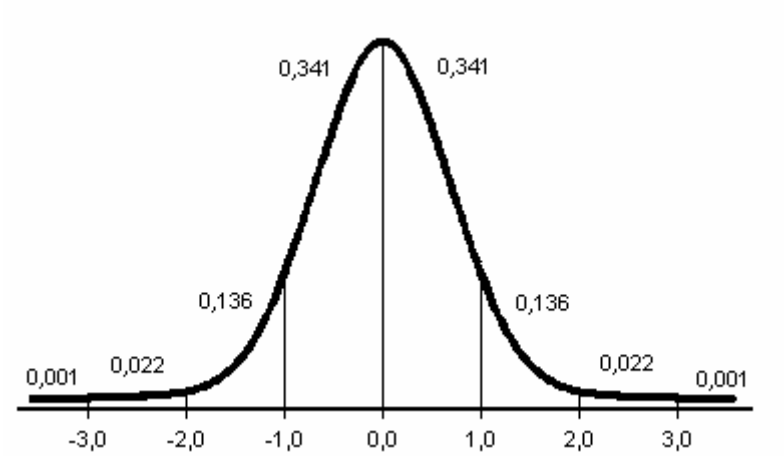


Figura 7.8 Curva normal tipificada y proporción de casos entre puntuaciones típicas

- Tiene forma de campana, por lo que los valores centrales son más probables que los extremos.
- Es simétrica respecto del centro de la curva, por lo cual la media, moda y mediana coinciden.
- Es asintótica respecto al eje de abcisas y su rango está entre -∞ y +∞.
- Tiene dos puntos de inflexión (cambio de curvatura) a una desviación típica a cada lado de la media.
- Cualquier combinación lineal de variables normalmente distribuidas también se distribuye normalmente.

De cara a los contrastes de hipótesis de estadísticos cuya distribución es la normal tipificada N(0,1) –media cero y desviación típica 1- es conveniente recordar que entre las puntuaciones típicas –1,96 y +1,96 se encuentra el 95% de los casos y entre las típicas –2 y +2 se encuentra el 95,5%; el 99% se encuentra entre las típicas -2,58 y +2,58. Posteriormente veremos que estos porcentajes coinciden con los niveles de confianza clásicos que se establecen en estadística inferencial para los contrastes de hipótesis.

8. Análisis Exploratorio

8.1 Introducción

Antes de proceder a cualquier análisis descriptivo de las variables objeto de estudio, es conveniente una exploración minuciosa de los datos, que permita identificar valores cuya lejanía de la parte central de la distribución puede alterar el resultado de los índices descriptivos. También es aconsejable ver si la distribución sigue o no pautas de normalidad, y en caso negativo ver procedimientos de transformación de los datos que permitan lograr dicha normalidad. Estos problemas y algunos más que comentaremos en este capítulo, pueden estudiarse mediante el procedimiento Explorar del SPSS.

8.2 Explorar

Este procedimiento (EXAMINE, en el lenguaje de sintaxis de SPSS) produce estadísticos descriptivos de resumen y representaciones gráficas de una variable tomada individualmente, o en función de otras variables de agrupamiento. Además de algunos de los estadísticos que se pueden obtener con los procedimientos **Frecuencias** y **Descriptivos**, **Explorar** aporta nuevos estadísticos, considerados resistentes, y representaciones gráficas de los datos ideadas por el creador de esta técnica de análisis, Tukey, y que publico en 1977 con el título *Exploratory Data Análisis*.

Para acceder al procedimiento, elegir

Analizar → Estadísticos descriptivos → Explorar...

y se muestra el cuadro de diálogo de la Figura 8.1

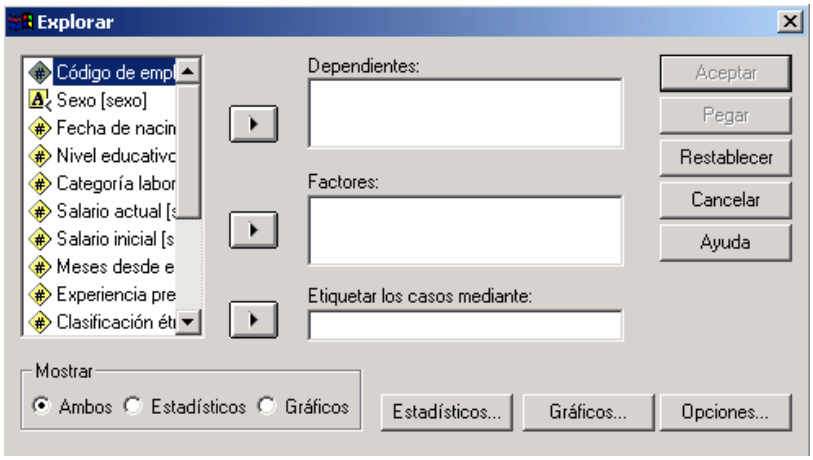


Figura 8.1 Cuadro de diálogo de Explorar

En la lista **Dependientes** se incorporan las variables dependientes que se pretenden analizar. En la lista **Factores** se incorporan las variables de agrupamiento si lo que se pretende es analizar las variables dependientes en función de los grupos de las variables Factores. En el cuadro **Etiquetar los casos**

mediante, se incorpora la variable que identifica los casos (si es que en el archivo hubiera alguna de este tipo), y en caso de no existir una variable así, el registro se identifica por el número de caso (recuerde el lector que el registro no siempre tiene el mismo número de caso, ya que si se ordena el archivo por alguna variable el número de caso del registro cambia). Por último, en el apartado **Mostrar** seleccionamos qué parte del análisis deseamos mostrar: si sólo los gráficos, sólo los estadísticos o ambos, que es la opción por defecto.

8.2.1 Estadísticos

Pulsando el botón Estadísticos del cuadro de dialogo Explorar, se accede al cuadro que se muestra en la Figura 8.2.

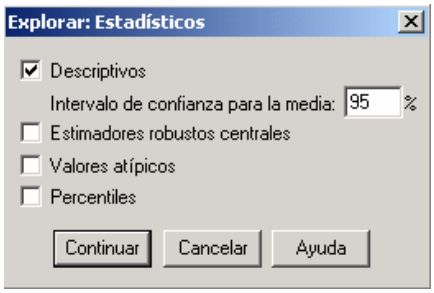


Figura 8.2 Cuadro de Estadísticos de Explorar

Los **estadísticos descriptivos** que incorpora explorar son la media aritmética, la mediana, la media recortada o trunciada al 5% (que es la media aritmética obtenida excluyendo el 5% de los casos con menor valor y el 5% de los casos con mayor valor), el error típico de la media, el intervalo de confianza para la media, la varianza, la desviación típica, los valores mínimo y máximo, la amplitud, el rango intercuartílico, los índices de asimetría y curtosis, y sus correspondientes errores típicos. Para el intervalo de confianza se puede elegir el nivel de confianza.

Respecto a los **Estimadores robustos centrales**, Explorar informa de 4 diferentes: el Estimador M de Huber, el Biponderado de Tukey, el Estimador M de Hampel, y la Onda de Andrews. Todos ellos no son más que medias ponderadas en donde los pesos asignados a los casos dependen de su distancia al centro de la distribución; los centrales reciben un peso de 1 y los demás valores van pesando menos a medida que se alejan del centro de la distribución. Con esta técnica, la media calculada no se ve tan afectada por los valores extremos de la distribución que tanto afectan a la media aritmética.

En los **valores atípicos** se muestran los 5 casos con valores más bajos y los 5 con valores más altos. Si existen empates en los valores ocupados por el quinto caso más pequeño o más grande, se indica en los resultados dicha circunstancia.

Los **percentiles** mostrados son el 5, 10, 25, 50, 75 , 90 y 95. El método de cálculo de estos percentiles es el HAVERAGE, por el cual se asigna al percentil buscado el valor que ocupa la posición $i = p(n+1)$ ordenados los casos de forma ascendente, donde p es la proporción correspondiente al percentil y n es el tamaño muestral. Si i no es entero, el valor del percentil se obtiene por interpolación. El SPSS incluye otros métodos de cálculo de percentiles, pero sólo pueden obtenerse mediante sintaxis.

El resultado de estos estadísticos para la variable **salario actual** (archivo *Datos de empleados)* como variable dependiente, el **sexo** como factor y la variable **código de empleado** como variable de identificación, es el mostrado en la Tabla 8.1.

Tabla 8.1 Resultados estadísticos del procedimiento Explorar

Descriptivos

		Salario actual			
		Sexo			
		Hombre		Mujer	
		Estadístico	Error típ.	Estadístico	Error típ.
Media		$41,441.78	$1,213.97	$26,031.92	$514.26
Intervalo de confianza para la media al 95%	Límite inferior	$39,051.19		$25,018.29	
	Límite superior	$43,832.37		$27,045.55	
Media recortada al 5%		$39,445.87		$25,248.30	
Mediana		$32,850.00		$24,300.00	
Varianza		380219336		57123688	
Desv. típ.		$19,499.21		$7,558.02	
Mínimo		$19,650		$15,750	
Máximo		$135,000		$58,125	
Rango		$115,350		$42,375	
Amplitud intercuartil		$22,675.00		$7,012.50	
Asimetría		1,639	,152	1,863	,166
Curtosis		2,780	,302	4,641	,330

Estimadores-M

	Salario actual			
Sexo	Estimador-M de Huber[a]	Biponderado de Tukey[b]	Estimador-M de Hampel[c]	Onda de Andrews[d]
Hombre	$34,820.15	$31,779.76	$34,020.57	$31,732.27
Mujer	$24,607.10	$24,014.73	$24,421.16	$24,004.51

a. La constante de ponderación es 1,339.

b. La constante de ponderación es 4,685.

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

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

Percentiles

		Sexo			
		Hombre		Mujer	
	Percentiles	Promedio ponderado(definición 1)	Bisagras de Tukey	Promedio ponderado(definición 1)	Bisagras de Tukey
Salario actual	5	$23,212.50		$16,950.00	
	10	$25,500.00		$18,660.00	
	25	$28,050.00	$28,050.00	$21,487.50	$21,525.00
	50	$32,850.00	$32,850.00	$24,300.00	$24,300.00
	75	$50,725.00	$50,550.00	$28,500.00	$28,500.00
	90	$69,325.00		$34,890.00	
	95	$81,312.50		$40,912.50	

Valores extremos

			Código de empleado		Valor	
			Sexo		Sexo	
			Hombre	Mujer	Hombre	Mujer
Salario actual	Mayores	1	29	371	$135,000	$58,125
		2	32	348	$110,625	$56,750
		3	18	468	$103,750	$55,750
		4	343	240	$103,500	$54,375
		5	446	72	$100,000	$54,000
	Menores	1	192	378	$19,650	$15,750
		2	258	338	$21,300	$15,900
		3	372	224	$21,300	$16,200
		4	22	411	$21,750	$16,200
		5	65	90	$21,900	$16,200

En la tabla de los Percentiles, además de los ya mencionados se muestran, las denominadas Bisagras de Tukey que son los percentiles 25, 50 y 75; sin embargo, algunos de los valores difieren porque el método de calculo es diferente al de los percentiles (en este caso se calculan con el método WAVERAGE).

8.2.2 Gráficos

Se pueden obtener varios tipos de gráficos: diagrama de caja, diagrama de tallo y hojas, histogramas, gráficos de normalidad y gráficos de dispersión. También se obtienen algunos estadísticos relacionados con los supuestos de normalidad y homogeneidad de varianzas. Para acceder a los gráficos se pulsa en el botón **Gráficos** del cuadro de diálogo **Explorar** y aparece el cuadro de diálogo de la Figura 8.3.

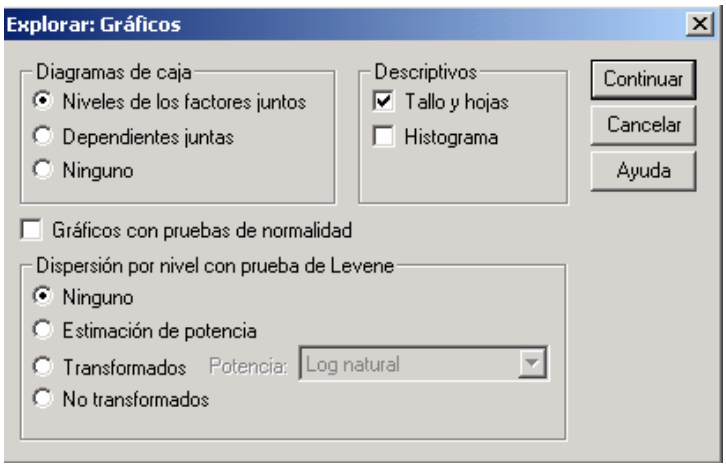


Figura 8.3. Cuadro de Gráficos de Explorar

8.2.2.1 Diagramas de caja

Mediante el diagrama de caja se puede visualizar algunos elementos relevantes de una distribución. El diagrama señala los 3 cuartiles de la distribución, y sobre el primero y tercero construye la caja, lo cual significa que en esa distancia se encuentra el 50% de las observaciones. Esto nos da un primer indicio gráfico de la dispersión de la muestra. También nos da una visión de la simetría, pues señala, en el interior de la caja, la mediana. Una mediana centrada en la caja es un indicio de

distribución simétrica –al menos en la parte central de la distribución. Estos tres valores se muestran en las tablas de los estadísticos de posición (los percentiles) con la denominación de *bisagras de Tukey*, ya mencionadas. Además, el diagrama señala los casos atípicos y los casos extremos. Los primeros están a 1,5 veces la distancia de la caja (el rango intercuartílico), desde los cuartiles uno y tres, y los extremos se encuentran a 3 veces la distancia de la caja desde esos mismos cuartiles. Las líneas que se dibujan desde la caja, van hasta los valores inferior y superior que no son atípicos.

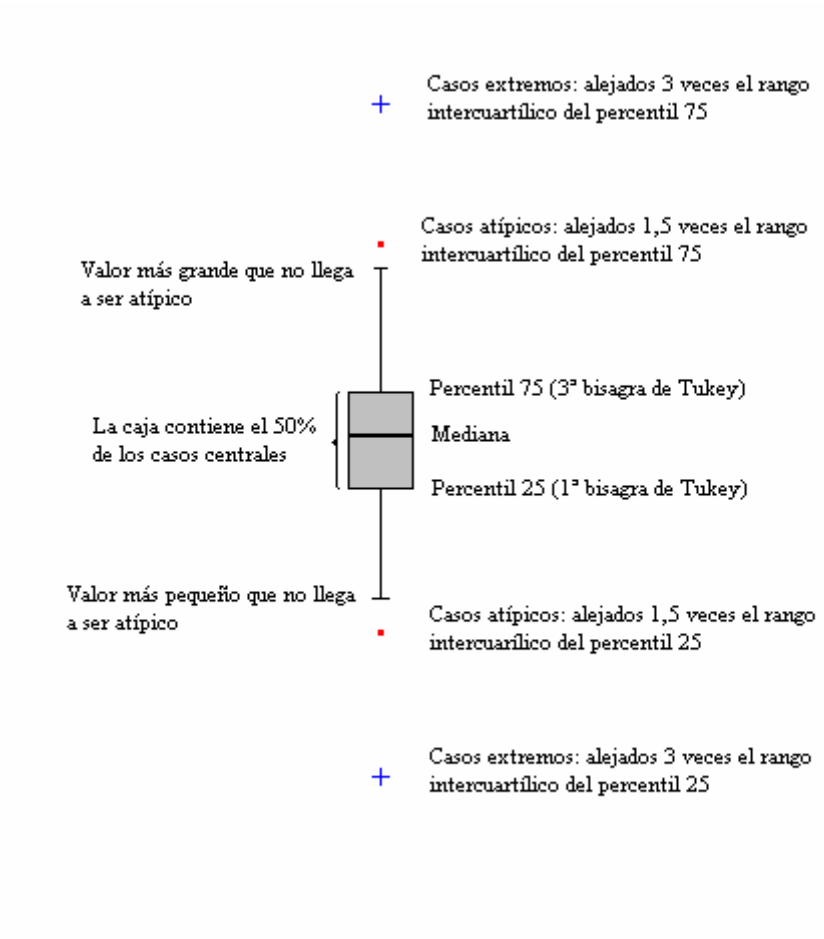


Figura 8.3 Detalles de un diagrama de caja

Entre las opciones del diagrama se encuentra el confeccionar un diagrama de caja de la variable dependiente para cada **nivel de la variable factor**. En este caso, si se han seleccionado varias variables dependientes, para cada una de ella se mostrará un gráfico distinto. También se puede elegir que se dibujen en el mismo gráfico todas las variables **dependiente juntas**. Por último, se puede optar por no realizar gráficos.

8.2.2.2 Diagrama de Tallo y hojas

Este diagrama es muy parecido a los histogramas, pero proporciona una información más precisa sobre la distribución de los casos. En la Figura 8.5 se muestra el diagrama de tallo y hojas de la variable **edad** del archivo *Datos de empleados* (esta no es una variable original del archivo, sino que se ha creado a

partir de la variable **fechnac**), sobre una muestra aleatoria del 35% de los casos del archivo. Del mismo modo que sucede en el histograma, la longitud de la línea refleja el número de casos que hay en cada intervalo. Cuando hay muchos, el tallo (en el caso de la variable edad el tallo son las decenas y las unidades son las hojas) puede ocupar más de una línea, e incluso, como es el caso, cada hoja representar a más de un caso, como sucede con nuestro diagrama. Si los tallos ocupan, por ejemplo, dos líneas, en la primera van desde el dígito de unidad 0 a 4 y en la segunda desde 5 a 9. En otras ocasiones, cada tallo puede ocupar hasta 5 líneas. Otra información esencial para entender el diagrama es el ancho del tallo (*stem width*). En nuestro diagrama este ancho es 10 lo que significa que el valor del tallo hay que multiplicarlo por 10.

```
 Frequency    Stem &  Leaf
26,00          3 .   111222223334&
58,00          3 .   55566677777777888888888999999
19,00          4 .   001122233
13,00          4 .   567779
14,00          5 .   11224&
15,00          5 .   567999&
14,00          6 .   023334
 7,00          6 .   78&
 5,00          7 .   02&
 Stem width:          10
 Each leaf:           2 case(s)
 & denotes fractional leaves.
```

Figura 8.5 Diagrama de tallo y hojas de la variable *edad*

Las hojas completan la información del tallo. Un tallo de 5 con una hoja de 1 representa una edad de 51 años. El número de casos que representa cada hojas también se muestra (*Each leaf*), y suele estar en función del tamaño muestral.

Cuando el ancho del tallo vale 10 los dígitos de las hojas son unidades; cuando vale 100 los dígitos de las hojas son decenas; cuando vale 1000 los dígitos de las hojas son centenas, y así sucesivamente. En la Figura 8.6 se muestra el diagrama del salario actual, y se ve que el ancho del tallo es de 10000,por lo que las hojas representan millares. Para esta muestra, cada hoja representa 3 casos.

```
Frequency     Stem &  Leaf
    33,00       1 .   56667789999
   110,00       2 .   000011111112222222223333344444444444
   115,00       2 .   555555556666666667777777778888889999999
    80,00       3 .   000000000001111112233333444
    32,00       3 .   55556677889
    20,00       4 .   0001233&
```

```
   12,00        4 .  5678&
   12,00        5 .  0124&
    7,00        5 .  556
   53,00 Extremes    (>=56750)
Stem width:     10000
Each leaf:       3 case(s)
& denotes fractional leaves.
```

Figura 8.6 Diagrama de tallo y hojas del *salario actual*

8.2.2.3 Histograma

El histograma es el diagrama que permite representar gráficamente datos de una variable cuantitativa continua. Se construye agrupando los datos en intervalos y levantando barras de altura proporcional a las frecuencias de cada intervalo. SPSS adapta de manera automática el número de intervalos a los datos, pero siempre es posible modificarlos en el Editor de gráficos. La Figura 8.7 muestra un histograma de la variable **salario actual** de modo que el lector pueda compararlo con el diagrama de tallo y hojas de la Figura 8.6. El histograma se ha girado (en el **Editor de gráficos**) para que la comparación sea más sencilla.

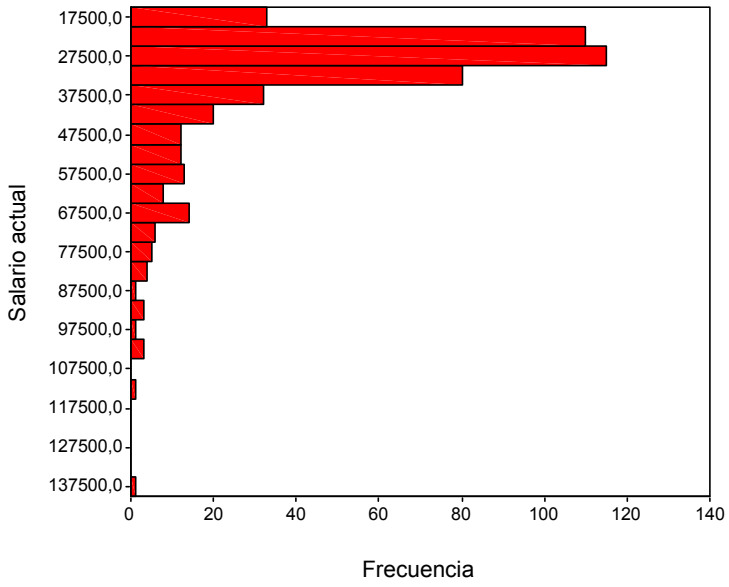


Figura 8.7 Histograma de la variable *salario actual*

8.3 Contraste de supuestos

La mayor parte de los procedimiento de análisis estadísticos denominados paramétricos, se basan en el cumplimiento, entre otros, de dos supuestos: normalidad de las distribuciones, y homocedasticidad u homogeneidad de las varianza. En el procedimiento Explorar se pueden contrastar estos dos supuestos tanto de forma gráfica como analítica.

8.3.1 Normalidad

Se pueden obtener dos tipos de gráficos de normalidad: uno con tendencia y otro sin tendencia, y dos pruebas de significación de la normalidad: *Kolmogorov-Smirnov* y *Shapiro-Wilk*. En general, se ofrece el estadístico de *Kolmogorov-Smirnov* (con las probabilidades de la prueba de *Lilliefors*) y además el de *Shapiro* cuando el tamaño muestral es menor o igual de 50. Para obtener los gráficos y las pruebas de significación, se señalan las opciones que se muestran en la Figura 8.8.

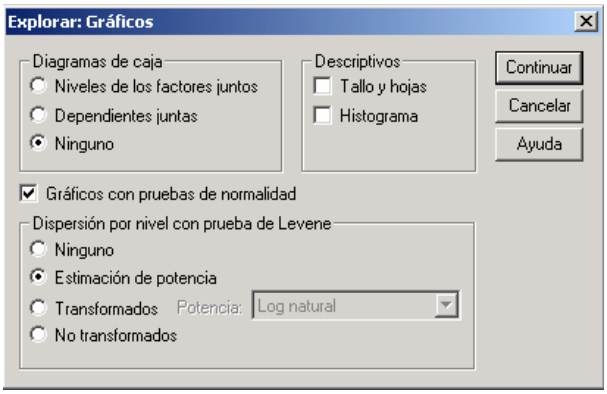


Figura 8.8 Opciones de gráficos de normalidad y de homogeneidad de varianza del cuadro Gráficos de Explorar.

Para la variable **salario actual** en función del **nivel de estudios** (variable categórica creada a partir de **educ**, y que tiene cuatro categorías –ver las categorías en la tabla de resultados de SPSS), las pruebas de significación correspondientes se muestran en la Tabla 8.2.

Tabla 8.2 Pruebas de significación de normalidad de la variable salario actual

Resumen del procesamiento de los casos

		Casos					
		Válidos		Perdidos		Total	
	ESTUDIO	N	Porcentaje	N	Porcentaje	N	Porcentaje
Salario actual	Primarios (8 años)	53	100,0%	0	,0%	53	100,0%
	Secundarios (12 años)	190	100,0%	0	,0%	190	100,0%
	Medios (de 14 a 16)	181	100,0%	0	,0%	181	100,0%
	Superiores (más de 16)	50	100,0%	0	,0%	50	100,0%

Pruebas de normalidad

	ESTUDIO	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
Salario actual	Primarios (8 años)	,119	53	,057			
	Secundarios (12 años)	,079	190	,006			
	Medios (de 14 a 16)	,154	181	,000			
	Superiores (más de 16)	,113	50	,148	,951	50	,076

a. Corrección de la significación de Lilliefors

De los cuatro grupos formados, las distribuciones del salario actual de quienes tienen estudios Secundarios y Medios no se distribuyen normalmente, es decir hay que rechazar la hipótesis de normalidad (Sig. < 0,05), y sí son normales las de quienes tienen estudios Primarios o Superiores (Sig. > 0,05).

Veamos ahora los gráficos de normalidad para el grupo de estudios Superiores (con distribución normal) y el grupo de estudios Medios (con distribución no normal). En la Figura 8.9 se observa el Gráfico Q-Q de normalidad con tendencia

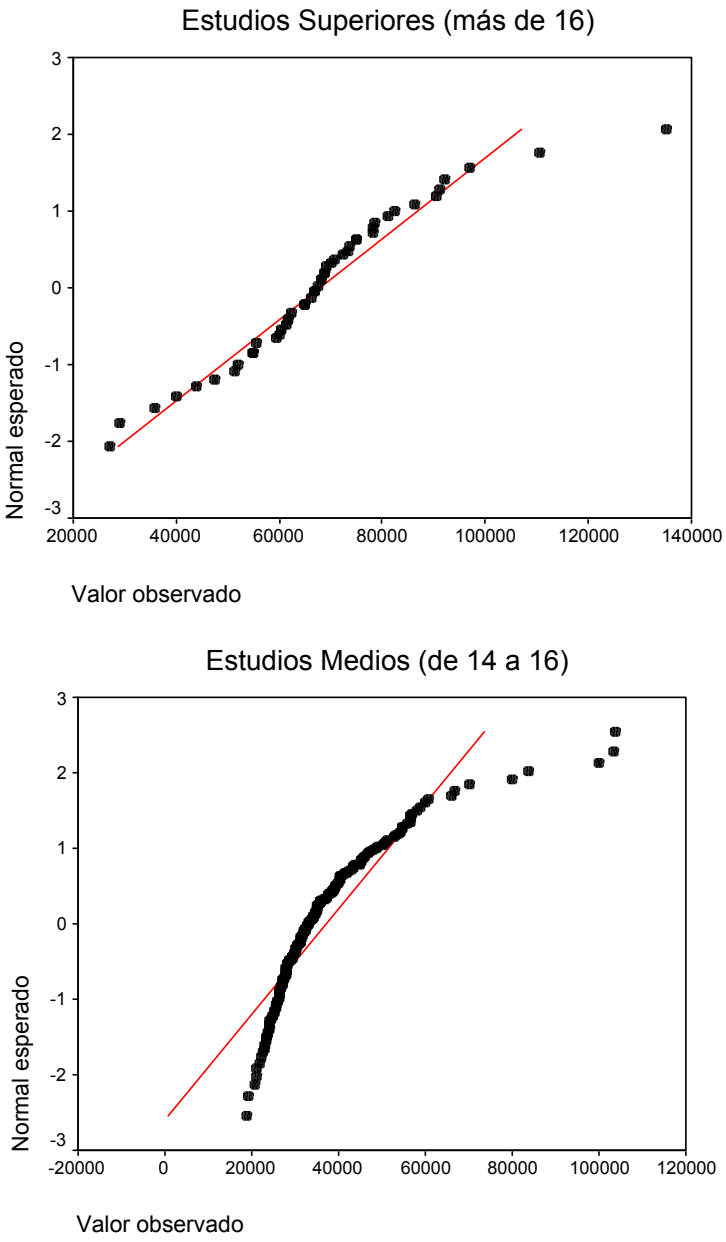


Figura 8.9 Gráficos Q-Q de normalidad con tendencias

En este tipo de gráficos, cada valor observado (eje de abcisas) es comparado con la puntuación típica que teóricamente le correspondería al valor en una distribución normal estandarizada o tipificada (eje de ordenadas). Las desviaciones de la diagonal, que representa la perfecta normalidad, representan desviaciones de la normalidad. Por otro lado, un gráfico *Q-Q normal sin tendencias*, muestra las diferencias entre la puntuación típica observada de cada valor, y su correspondiente puntuación típica normalizada. En el eje de abcisas está el valor observado y en el de ordenadas el tamaño de las diferencias entre las puntuaciones típicas observadas y las esperadas en caso de normalidad. Cuando la distribución es normal, los puntos del gráfico se distribuyen de manera aleatoria en

torno al valor 0 del eje de ordenadas, mientras que si no es normal, los puntos mostrarán alguna tendencia o forma más o menos estructurada. En las gráficas *Q-Q normal sin tendencias* de la Figura 8.10 se observa esta circunstancia, para el grupo con estudios primarios (distribución normal) y el grupo de estudios medios (distribución no normal).

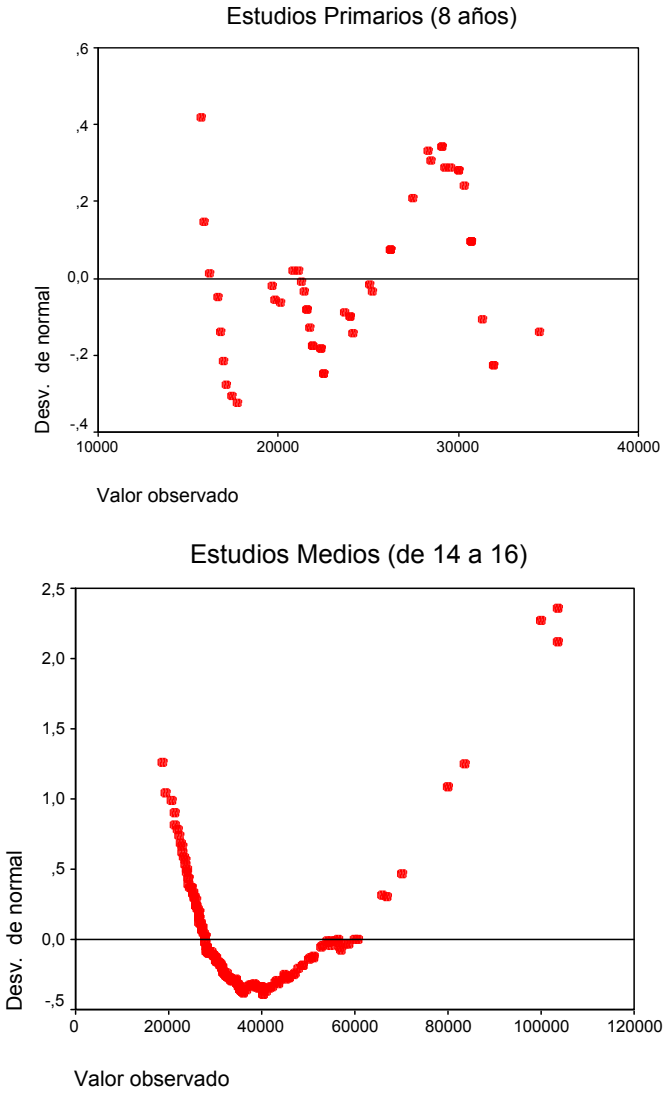


Figura 8.10 Gráficos Q-Q normal sin tendencia

8.3.2 Homogeneidad de varianzas

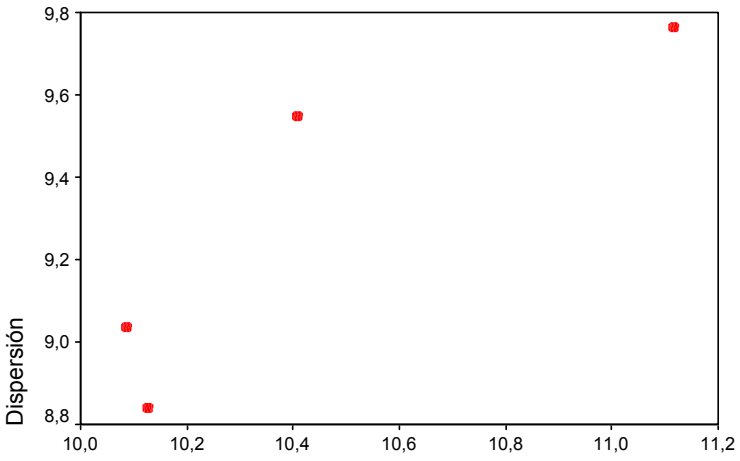
Como ya se ha señalado, el procedimiento ***Explorar*** también informa de si las varianzas son o no homogéneas. En el caso del salario actual se puede ver que no hay homocedasticidad entre los grupo de estudio, tal como informa el Estadístico de *Levene* que se puede ver en la Tabla 8.3

Tabla 8.3 Prueba de significación de homogeneidad de varianzas

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
Salario actual	Basándose en la media	28,085	3	470	,000
	Basándose en la mediana.	21,799	3	470	,000
	Basándose en la mediana y con gl corregido	21,799	3	266,893	,000
	Basándose en la media recortada	24,767	3	470	,000

En el gráfico de dispersión por nivel que se muestra en la Figura 8.10 se ve que las varianzas son muy distintas entre los grupos, ya que los puntos del gráfico no se encuentran alineados en sentido horizontal. Como informa el gráfico, los ejes están construidos sobre el logaritmo neperiano de la dispersión frente al logaritmo neperiano del nivel (es decir del promedio de salario por nivel).



* Gráfico de LN de dispersión por LN de nivel

Inclinación = ,797 Potencia para transformación = ,203

Figura 8.10. Gráfico de dispersión por nivel de salario actual según nivel de estudios

El gráfico muestra el valor de la pendiente de la línea de regresión que ajusta los puntos, y a partir de este valor ofrece una estimación de la potencia a la que habría que elevar las puntuaciones de la variable dependiente para intentar homogeneizar la varianzas de la variable en cada nivel del factor, aunque no siempre se consigue. En este caso, el valor de la potencia es 0,203 (resultado de restar a 1 el valor de la pendiente: 1 – 0,797 = 0,203). No obstante, lo habitual es utilizar potencias redondeadas a múltiplos de 0,5. Por último, señalaremos que las potencias más

comúnmente utilizadas para transformar datos son las siguientes: -1 = recíproco; -1/2 = recíproco de la raíz cuadrada; Logaritmo neperiano; raíz cuadrada; el cuadrado; el cubo. Estas son las transformaciones que contiene el SPSS.

9. Análisis de datos categóricos

9.1 Introducción

En el ámbito de las ciencias sociales es habitual el estudio de variables con una escala de medida nominal u ordinal con pocas categorías. Pensemos, por ejemplo, en el estado civil, la clase social, el sexo, religión que se profesa, tratamientos aplicados en determinados síndromes, grado de satisfacción ante determinado producto, etc. Para este tipo de datos, SPSS dispone de un procedimiento, denominado Tablas de contingencia, que permite el análisis estadístico para determinar si las variables están relacionadas o por el contrario son independientes. Aunque este procedimiento permite incorporar múltiples variables, sólo nos vamos a centrar en el análisis que se refiere a dos variables, es decir sólo vamos a tratar con tablas de contingencia de doble entrada[5].

9.2 Tablas de contingencia

Como se ha señalado, los datos categóricos se disponen en tablas de doble entrada. Como ejemplo, tomemos una tabla de contingencia de dos dimensiones en donde se cruzan la **ideología política** y la **opinión ante el aborto**, datos que se muestran en la Tabla 9.1.

Tabla 9.1 Tabla de contingencia de ideología política y opinión ante el aborto

Recuento

		Opinión ante el aborto			
		A favor	Sin opinión	En contra	Total
Ideología política	Derecha	8	9	25	42
	Centro	18	6	15	39
	Izquierda	28	3	8	39
Total		54	18	48	120

Para acceder al procedimiento ***Tablas de contingencia*** hay que seguir la secuencia:

Analizar → Estadísticos descriptivos → Tablas de contingencia

y se muestra el cuadro de diálogo de la Figura 9.1

[5] Para el alumno que desee ampliar sus conocimientos sobre el análisis sobre datos categóricos cuando hay más de dos variables, recomendamos el seguimiento del curso "Análisis de Datos Categóricos", que imparte el Dr. Antonio Pardo, de la UAM, y forma parte del elenco de cursos ofertados por el Programa de Doctorado de Metodología de las Ciencias del Comportamiento.

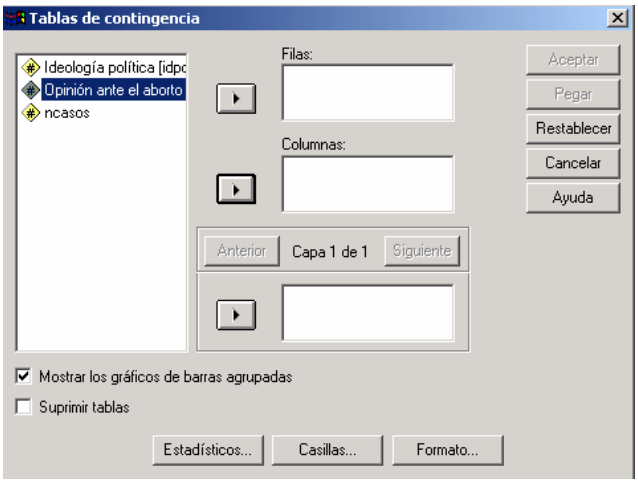


Figura 9.1 Cuadro de diálogo de Tablas de contingencia

En este cuadro se selecciona la variable que aparecerá en las filas, la que aparecerá en las columnas, y si se quiere cruzar este par de variables con otra variable de agrupamiento, trasladaríamos ésta a la lista de la Capa. Además, podemos determinar si se muestra el gráfico de barras agrupadas y si se suprime la tabla (por defecto, se muestra la tabla y no el gráfico). Si para las variables ideología política y opinión ante el aborto marcamos la opción **Mostrar gráficos de barras agrupadas** el resultado es el que se muestra en la Figura 9.2.

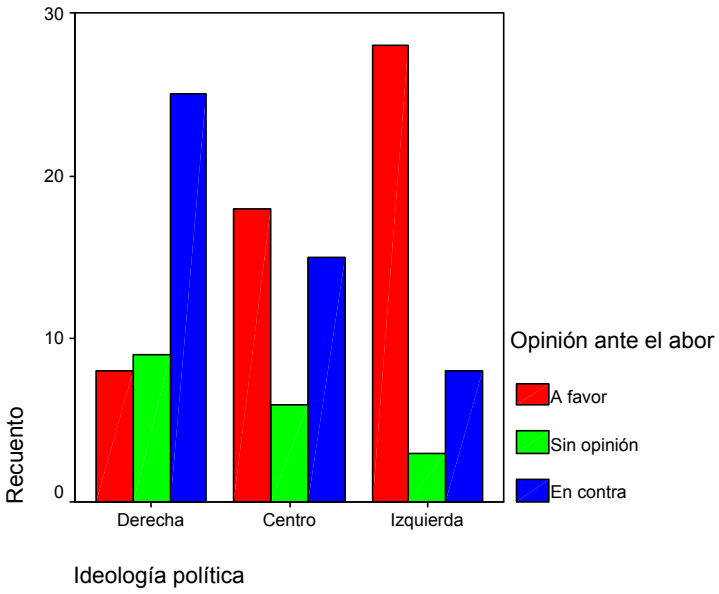


Figura 9.2 Gráfico de barras agrupadas de *ideología pol´tica* y *opinión ante el aborto*

9.3 Estadísticos

Para determinar si hay relación o independencia entre las variables no basta observar la tabla, sino que es preciso llevar a cabo una prueba de significación. Estas pruebas se encuentran y seleccionan en el cuadro de diálogo que se muestra al pulsar el botón **Estadísticos** (Figura 9.3)

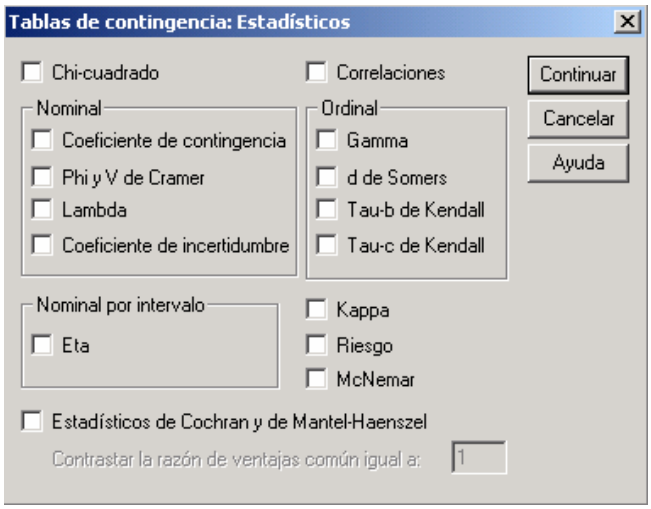


Figura 9.3 Cuadro de estadísticos de Tablas de contingencia

9.3.1 Chi-cuadrado

El más familiar de estos estadísticos es Chi-cuadrado, propuesto por **Pearson**, que contrasta la hipótesis de que los dos criterios de clasificación empleados son independientes. Para ello compara las frecuencias observadas con las esperadas en el caso de que efectivamente fueran independientes. El cálculo de las frecuencias esperadas es muy sencillo y su fundamento es el siguiente, tomando como ejemplo los datos de la Tabla 9.1 que nuevamente mostramos.

Recuento

		Opinión ante el aborto			
		A favor	Sin opinión	En contra	Total
Ideología política	Derecha	8	9	25	42
	Centro	18	6	15	39
	Izquierda	28	3	8	39
Total		54	18	48	120

La proporción, por ejemplo, de personas que se reconocen con ideología de derechas respecto del total de la muestra, es 42/120 = 0,35. Si ambos criterios de clasificación fueran independientes este porcentaje debería ser el mismo para cada categoría de la variable opinión ante el aborto: el 35% de los sujetos que están a favor del aborto sería de derechas (0,35 x 54 = 18,9 sujetos); el 35% de los que no tienen opinión sería de derecha (0,35 x 18 = 6,3), y el 35% de los que están en contra sería de derecha (0,35 x 48 = 16,8). Es decir, para obtener las frecuencias esperadas basta con multiplicar las respectivas frecuencias marginales y dividir por el número total de casos de la muestra.

Obtenidas así las frecuencias esperadas, el estadístico χ^2 o chi-cuadrado, se calcula de la siguiente forma:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \sum_i \sum_j \frac{n_{ij}^2}{m_{ij}} - n$$

donde n_{ij} son las frecuencias empíricas u observadas, y m_{ij} son las frecuencias esperadas. Este estadístico sigue el modelo de probabilidad χ^2 con los grados de libertad resultantes de multiplicar el número de categorías de las filas menos uno por el número de categorías de las columnas menos uno; en este caso habría (3-1)x(3-1)=4 grados de libertad. La tabla que muestra el valor del estadístico y su significación es la que se muestra en la Tabla 9.2.

Tabla 9.2 Pruebas de chi-cuadrado de Tablas de contingencia

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	22,798[a]	4	,000
Razón de verosimilitud	24,046	4	,000
N de casos válidos	120		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5,85.

El valor del estadístico es 22,798, y su probabilidad asociada es inferior a 0,05 y por tanto es preciso rechazar la hipótesis de independencia de ambas variables. Además de este estadístico, también se muestra otro estadístico denominado **Razón de verosimilitud** que se obtiene mediante

$$\textbf{Razón de verosimilitud} = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right)$$

que es un estadístico que también se distribuye según χ^2 (e igual que el anterior se interpreta) y que se utiliza para estudiar la relación entre variables categóricas en los modelos log-lineales.

Cuando la tabla de contingencia es de dos variables dicotómicas, los resultados incluyen información adicional: la corrección por continuidad de **Yates** y el estadístico exacto de **Fisher**. El primero consiste en restar 0,5 al valor absoluto de las diferencias $n_{ij} - m_{ij}$ del estadístico χ^2, antes de elevarlas al cuadrado. Por su parte el estadístico exacto de Fisher ofrece la probabilidad exacta de obtener las frecuencias de hecho obtenidas o cualquier otra combinación alejada de la hipótesis de independencia; este estadístico está basado en la distribución hipergeométrica y en la hipótesis de independencia.

9.3.2 Correlaciones

Esta opción muestra dos tipos de correlaciones: el coeficiente de correlación de Pearson, y el de Spearman. El de Pearson es una medida de asociación lineal entre variables cuantitativas (variables de escala en la denominación de SPSS), mientras que el de Spearman es también una medida de asociación lineal para variables ordinales (ambos índices se tratan en el capítulo 14). Estos coeficientes son poco útiles para estudiar las pautas de relación entre variables categóricas.

9.3.3 Datos nominales

El estadístico χ^2 contrasta la hipótesis de independencia pero no indica la fuerza de esa asociación. Su valor depende no sólo de las diferencias entre frecuencias observadas y esperadas, también depende del número de casos de las muestra. Para tamaños muestrales grandes, diferencias pequeñas entre frecuencias observadas y esperadas pueden dar lugar a valores de χ^2 grandes y por tanto significativos. Por ello hay otras medidas de asociación, cada una con diferencias entre sí en como son afectadas por las diferentes características de la distribución de las variables. En cualquier caso estas medidas sólo informan del grado de asociación, no de la naturaleza de la misma.

9.3.3.1 Medidas basadas en chi-cuadrado

Son medidas que ofrecen valores entre 0 y 1, e intentan minimizar el efecto que el tamaño de la muestra tiene sobre χ^2.

- **Coeficiente de contingencia C**: Su fórmula es $C = \sqrt{\chi^2 / (\chi^2 + n)}$, toma valores entre 0 y 1, pero difícilmente llega a 1. Su valor máximo depende del número de filas y columnas. Si ambos valores coinciden (k) el valor máximo de C es: $C_{máx.} = \sqrt{(k-1)/k}$. Un valor 0 indica independencia y un valor cercano a 1 indica asociación.

- **Phi**. El coeficiente *Phi* se obtiene de las siguiente forma: $\Phi = \sqrt{\chi^2 / n}$. En la tablas con dos variables dicotómicas, Φ toma valores entre 0 y 1. En tablas en las que una variable tiene más de dos categorías, Φ puede tomar valores mayores que 1, pues χ^2 puede ser mayor que el tamaño muestral.

- ***V* de *Cramer*** es ligeramente diferente a *Ph,* $V_{Cramer} = \sqrt{\chi^2 / [n(k-1)]}$, siendo k el menor del número de filas y de columnas. Este índice nunca excede de 1 y en tablas 2x2 toma el mismo valor que *Phi*.

9.3.3.2 Medidas basadas en la reducción proporcional del error (RPE)

Con este tipo de medidas se consigue reducir la probabilidad de cometer un error de predicción cuando en vez de predecir un caso o grupo de casos como perteneciente a alguna categoría de una variable, se utiliza también la información de las probabilidades de esa variable en cada categoría de la otra.

- **Lambda**. Esta opción proporciona dos tipos de medidas de asociación, la lambda y la tau, desarrolladas por Goodman y Kruskal (1979).

  El índice **Lambda** expresa la proporción de error de predicción que conseguimos reducir al predecir la clasificación de un caso o grupo de casos como perteneciente a una categoría de una variable utilizando, además de la información de esta variable, la información de la variable con la que está cruzada.

Si tomamos como referencia los datos de la Tabla 9.1 al estimar a que grupo de opinión ante el aborto pertenece un sujeto cualquiera diríamos que al de que están "A favor" pues la probabilidad será de 54/120 = 0,45, frente a una probabilidad del (18+48)/120 = 0,55 de que la predicción esté equivocada. Si además tenemos en cuenta la variable ideología política, clasificando a los de derecha en el grupo de "en contra" cometemos una error de (8+9)/120 = 0,1417, a los de centro en el grupo de "a favor" cometemos un error de (6+15)/120 = 0,175, y a los de izquierda en el grupo de "a favor" cometeremos un error de (3+8)/120 = 0,0917. Procediendo así, cometemos un error de 0,1417+0,175+0,0917 = 0,4083, y por tanto hemos reducido la probabilidad de error de 0,55 a 0,4083, es decir 0,1417. Esta reducción respecto del error de predicción inicial al considerar sólo la variable opinión ante el aborto, representa una proporción de 0,1417/0,55 = 0,258 que es el valor que se muestra en la Tabla 9.3

Tabla 9.3 Medidas de asociación basadas en la RPE de Tablas de contingencia

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Lambda	Simétrica	,257	,059	4,313	,000
	Ideología política dependiente	,256	,066	3,499	,000
	Opinión ante el aborto dependiente	,258	,075	3,074	,002
Tau de Goodman y Kruskal	Ideología política dependiente	,096	,036		,000[c]
	Opinión ante el aborto dependiente	,123	,046		,000[c]
Coeficiente de incertidumbre	Simétrica	,095	,036	2,633	,000[d]
	Ideología política dependiente	,091	,035	2,633	,000[d]
	Opinión ante el aborto dependiente	,099	,038	2,633	,000[d]

a. Asumiendo la hipótesis alternativa.

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

c. Basado en la aproximación chi-cuadrado.

d. Probabilidad del chi-cuadrado de la razón de verosimilitud.

Lambda toma valores entre 0 y 1. El valor 0 indica que la variable independiente no aporta nada en la reducción del error de predicción y un valor 1 indica que el error de predicción se ha conseguido reducir por completo.

Lambda tiene tres versiones: dos asimétricas, según que alguna de las variables se considere dependiente y la otra independiente, y una simétrica, para cuando no haya razón para distinguir las variables en dependiente o independiente. La expresión para la versión asimétrica es:

$$\lambda_{Y/X} = \frac{\sum_i \text{máx}_i(n_{ij}) - \text{máx}(n_{+j})}{n - \text{máx}(n_{+j})}$$

donde:

$máx_i(n_{ij})$ = la mayor de las frecuencias de la fila i

$máx(n_{+j})$ = la mayor de las frecuencias marginales de las columnas.

La fórmula de la otra versión asimétrica será:

$$\lambda_{X/Y} = \frac{\sum_j máx_j(n_{ij}) - máx(n_{i+})}{n - máx(n_{i+})}$$

donde:

$máx_j(n_{ij})$ = la mayor de las frecuencias de la columna j

$máx(n_{i+})$ = la mayor de las frecuencias marginales de las filas.

La versión simétrica se obtiene promediando el valor de las dos versiones asimétricas.

- El índice ***tau*** es similar a *lambda* pero con una lógica diferente. Al pronosticar a qué categoría de **opinión ante el aborto** pertenece, diremos que el 100(54/120) = 45% estarán "a favor", el 15% "sin opinión" y el 40% "en contra", estaremos clasificando de forma correcta una proporción de 0,385 (el promedio ponderado de las proporciones correctas de clasificación para cada categoría), y por tanto estaremos cometiendo un error de 1– 0,385 = 0,615. Si tenemos en cuenta la variable **ideología política**, y vamos clasificando por ideologías, el 100(8/42) = 19% de derecha estará "a favor", el 100(9/42) = 21,4% "sin opinión", y así sucesivamente, clasificaríamos a todos los sujetos de las tres ideologías. Al final, promediando ponderadamente, clasificaremos de forma correcta con una probabilidad de 0,4609 y por tanto cometeremos un error de 0,5391; es decir, reducimos la probabilidad de error de 0,615 a 0,5391, una diferencia de 0,0759, que representa una proporción de reducción de error respecto a la primera clasificación de 0,0759/0,615 = 0,123, que es el valor de la tau que se muestra en la Tabla 9.3 cuando la **opinión ante el aborto** se toma como variable dependiente.

Al igual que lambda, tau toma valores entre 0 y 1 y el significado de estos valores es el mismo.

La fórmula para obtener el valor de *tau* es la siguiente:

$$\tau_{Y/X} = \frac{n\sum_i\sum_j\left(n_{ij}^2 / n_{i+}\right) - \sum_j n_{+j}^2}{n^2 - \sum_j n_{+j}^2}$$

Hay dos versiones asimétrticas de *tau*. Para obtener el valor de $\zeta_{X/Y}$, sólo hay que intercambiar los papeles de X_i e Y_j en la expresión anterior.

- **Coeficiente de incertidumbre**. Elaborado por Theil (1970) este índice expresa el grado de incertidumbre que conseguimos reducir cuando utilizamos una variable para efectuar pronósticos sobre otra. Igual que lambda, posee dos versiones asimétricas y una simétrica, cuando no hay razón para tomar una u otra variables como dependiente o independiente. El índice se obtiene a partir de:

$$I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)}$$

donde:

$$I(X) = -\sum_i \left[\left(n_i / n \right) \ln \left(n_i / n \right) \right]$$

$$I(Y) = -\sum_j \left[\left(n_j / n \right) \ln \left(n_j / n \right) \right]$$

$$I(XY) = -\sum_i \sum_j \left[\left(n_{ij} / n \right) \ln \left(n_{ij} / n \right) \right]$$

n_i = frecuencias marginales filas

n_j = frecuencias marginales columnas

n_{ij} = frecuencias conjuntas ($n_{ij}>0$)

Para obtener $I_{X/Y}$ basta con intercambiar I(X) por I(Y). Para la versión simétrica, hay que multiplicar $I_{X/Y}$ por 2 después de añadir I(X) al denominador.

En la tabla de resultados (Tabla 9.3), además del valor de estos índices también se muestra el error típico asintótico, que es el error cuando no se supone la independencia entre variables, el valor del estadístico T, y su significación estadística. También se muestra, al pie de la tabla una serie de notas aclaratorias sobre determinados aspectos y condiciones de cómo se han hecho algunos cálculos.

9.3.4 Datos ordinales

Para datos ordinales, ***Tablas de contingencia*** calcula una serie de estadísticos, basados todos ello en el mismo principio: el concepto de inversión y no inversión en los órdenes de los datos. Esto quiere decir que si dos valores de un caso cualquiera son mayores o menores que los de otro caso, se dice que no hay inversión, pero si el valor de un caso en una variables es mayor que el de otro en esa misma variable, pero menor en la otra, se dice que hay una inversión en el orden. Se designa la *no inversión* como ***P*** y la *inversión* como ***Q***. Si dos casos tiene valores idénticos en una o en las dos variables se dice que hay un *empate* (***E***). Las medidas para estos datos se diferencian entre sí en el tratamiento de a los empates.

- **Gamma**. La fórmula es $\gamma = \left(n_P - n_Q \right) / \left(n_P + n_Q \right)$. Si la relación entre las variables es perfecta y positiva todas las comparaciones serán no inversiones, y el valor será 1; si, al contrario, la relación es perfecta y negativa, todo serán inversiones y el valor será -1. Por último, tendrá un valor 0 cuando el número de no inversiones e inversiones sea el mismo.

- **d de Somers**. Este índice es para cuando una variable se considera independiente y otra dependiente. Su fórmula es: $d = (n_P - n_Q) / (n_P + n_Q + n_{E(Y)})$, siendo $n_{E(Y)}$, el número de pares empatados en la variable dependiente.
- **Tau-b de Kendall**. Su fórmula es $\tau_b = (n_P - n_Q) / \sqrt{(n_P + n_Q + n_{E(X)})(n_P + n_Q + n_{E(Y)})}$ y sólo toma valores –1 y +1 en las tablas 2 x 2 sin frecuencias marginales con valor cero.
- **Tau-c de Kendall**. Su fórmula es: $\tau_c = 2m(n_P - n_Q) / [n^2(m-1)]$, siendo *m* el menor valor del número de filas y de columnas. Tau-c toma valores entre –1 y +1.

9.3.5 Nominal por intervalo

El coeficiente eta cuantifica el grado de asociación entre una variable cuantitativa y otra nominal. Su principal utilidad es que es un coeficiente que no supone linealidad (a diferencia de el de Pearson) y el cuadrado se puede interpretar como la proporción de varianza de la variable cuantitativa que es explicada por la nominal.

9.3.6 Índice de acuerdo Kappa

Este índice, elaborado por Cohen (1960) evalúa el acuerdo existente entre las clasificaciones de dos jueces diferentes sobre la misma muestra de sujetos. Para ilustrar el cálculo del índice pensemos en dos jueces que tienen que clasificar a 160 sujetos en cuatro categorías diferentes A, B, C ó D. La Tabla 9.4 muestra la clasificación conjunta.

Tabla 9.4 Clasificación conjunta de dos jueces en una muestra de 160 casos

Recuento

		Juez B				
		A	B	C	D	Total
Juez A	A	15	8	5	7	35
	B	10	20	8	5	43
	C	10	12	16	9	47
	D	7	7	9	12	35
Total		42	47	38	33	160

El número de coincidencias de ambos jueces, 63, es la suma de las frecuencias de la diagonal principal; esto representa una proporción de acuerdo de 63/160 = 0,3937. Por azar, esperaríamos una acuerdo igual a 40,2/160 = 0,2512, siendo 40,2 la suma de las frecuencias esperadas de la diagonal principal. Es decir, por azar la proporción de acuerdo sería de 0,2512 y por tanto la proporción de acuerdo máximo no debido al azar sería de 1–0,2512 = 0,7488. El índice capa es el cociente entre la diferencia entre el acuerdo observado y el esperado por azar y el acuerdo máximo posible descontado el azar, es decir (0,3937-0,2512)/0,7488 = 0,1903, que es el valor que se muestra en la Tabla 9.5 de resultados.

Tabla 9.5 Índice de acuerdo *kappa*

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Medida de acuerdo	Kappa	,190	,051	4,178	,000
N de casos válidos		160			

a. Asumiendo la hipótesis alternativa.

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

Este valor se interpreta teniendo en cuenta que el índice toma valores entre 0 (acuerdo nulo) y 1 (acuerdo total).

9.3.7 Índices de riesgo

En las tablas que hemos utilizado en los índices anteriores, los datos están tomados en el mismo momento temporal. No obstante, hay otra manera que consiste en hacer un seguimiento de una muestra de sujetos a lo largo de un período de tiempo. Este tipo de estudios, puede hacerse hacia delante o hacia atrás. Los primeros se conocen como diseños prospectivos o de *cohortes*, y los segundos como diseños *retrospectivos* o de *caso-control*.

Para este tipo de diseños longitudinales, para el caso de dos variables dicotómicas, los índices de riesgo nos proporciona una medida del riesgo relativo de un grupo de sujetos respecto de otro en función de la condición a la que pertenecen en otra variable. Pensemos, por ejemplo, en el la relación que pueda haber entre el hábito de fumar y padecer o no problemas vasculares. En la Tabla 9.6 se muestran los datos de una muestra de 190 sujetos.

Tabla 9.6 Tabla de contingencia de tabaquismo y problemas vasculares

Recuento

		Problemas vasculares		Total
		Sin problemas	Con problemas	
Tabaquismo	No fuman	30	50	80
	Fuman	12	98	110
Total		42	148	190

Entre los fumadores la proporción de problemas vasculares es 30/80 = 0,375, mientras que en los no fumadores es 12/110 = 0,109. El *riesgo relativo* será el cociente entre ambas proporciones 0,375/0,109 = 3,4375, y nos informa del número de veces que es más probable tener problemas vasculares de los sujetos que fuman respecto de los que no fuman. El valor 1 de este riesgo relativo significaría que no hay diferencias entre una condición y otra. Este sería un ejemplo de estudio de cohortes, en el que se mide el riesgo futuro debida a la presencia o ausencia de alguna condición.

En el diseño de caso-control, se busca hacia atrás la presencia o ausencia de algún factor desencadenante. Para estos mismos datos, se podría formar dos grupos en función de si tienen o no problemas vasculares y buscar su historia de tabaquismo. Así se puede calcular la razón (*ratio*) entre fumadores/no fumadores

tanto en el grupo con problemas como en el grupo sin problemas, y el cociente entre ambas *ratios* será considerado como un índice de riesgo relativo.

Con los datos de la Tabla 9.6 la *ratio* fumadores/no fumadores en el grupo de sujetos sin problemas vasculares es 30/12 = 2,5 y en el grupo de sujetos con problemas es 50/98 = 0,51. Por tanto el índice de riesgo será 2,5/0,51 = 4,9. Este valor se interpreta de la misma forma que el riesgo relativo, pero también se puede interpretar como que entre los sujetos que tiene problemas vasculares, es 4,9 veces más probable encontrar fumadores que no fumadores. En la Tabla 9.7 se muestran estos resultados.

Tabla 9.7 Índice de riesgo del procedimiento Tablas de contingencia

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Razón de las ventajas para Tabaquismo (No fuman / Fuman)	4,900	2,312	10,385
Para la cohorte Problemas vasculares = Sin problemas	3,438	1,878	6,291
Para la cohorte Problemas vasculares = Con problemas	,702	,585	,841
N de casos válidos	190		

En las dos últimas filas de la tabla con el índice de riesgo se muestran los límites inferior y superior del intervalo de confianza. Si entre estos límites se encuentra el valor 1 significa que el riesgo es el mismo en esa cohorte sea cual sea el supuesto factor de riesgo.

9.3.8 Proporciones relacionados. Índice de McNemar

Otro enfoque de los datos categóricos es en un diseño longitudinal del tipo antes después, determinar si ha habido cambio o no respecto de una cuestión concreta. La situación podría ser la de sondear a un grupo de sujetos sobre un asunto cualquiera, aplicarles algún tipo de tratamiento, y volver a sondearles después de este tratamiento.

Pensemos por ejemplo en los datos de la Tabla 9.8 con datos de un grupo de 190 sujetos a los que se les ha pedido opinión sobre una determinada cuestión antes y después de visionar un documental sobre dicha cuestión.

Tabla 9.8 Opinión sobre un asunto antes y después y probabilidad del estadístico de *McNemar*

Recuento

		Opinión después		
		En contra	A favor	Total
Opinión antes	En contra	60	45	105
	A favor	15	70	85
Total		75	115	190

	Valor	Sig. exacta (bilateral)
Prueba de McNemar		,000[a]
N de casos válidos	190	

a. Utilizada la distribución binomial

El estadístico de *McNemar* compara los cambios que se producen antes y después en ambas direcciones y determina la probabilidad de encontrar ese número concreto si las proporciones antes-después fueran iguales. De acuerdo con la hipótesis nula la proporción de cambios a favor-en contra debe ser la misma que la proporción del cambio en contra-a favor. Si el número de cambios no es muy grande SPSS intenta calcular la probabilidad exacta de encontrar un número tal de cambios, y para ello se basa en la distribución binomial con parámetros *n = número de cambios* y $\pi = 0{,}5$. Si el número de cambios es muy grande, SPSS ofrece una probabilidad aproximada basada en el estadístico de McNemar (1947) y en la distribución *ji-cuadrado*. Este estadístico se calcula de la siguiente manera:

$$X^2_{\text{McNemar}} = \frac{(\text{nº de cambios en una dirección - nº de cambios en la otra dirección})^2}{\text{nº total de cambios}}$$

Este estadístico se distribuye según *ji-cuadrado* con 1 grado de libertad. Para los datos de la Tabla 9.8 su valor sería.

$$X^2_{\text{McNemar}} = \frac{(45-15)^2}{45+15} = 15$$

En la tabla inferior de la Tabla 9.8 se muestra la probabilidad de un número de cambios como el de la tabla superior utilizando la distribución binomial. Por ello no se muestra el valor del estadístico de McNemar, que sólo se calcula cuando el tamaño muestral no supone un problema en la computación de los datos.

9.3.9 La prueba de Cochran y Mantel-Haenszel

Esta prueba se emplea en tablas 2 x 2 de diseños de *cohortes* o de *caso-control* cuando interviene una tercera variable, situación en la cual el estadístico *chi-cuadrdo* de **Pearson** sobre todos los datos agrupados puede dar resultados equívocos. Lo que se hace es analizar la muestra por estratos. Los estadísticos de **Cochran** y **Mantel-Haenszel** cotrastan la hipótesis de independencia condicional, es decir la hipótesis entre la variable dependiente (por ejemplo, problemas vasculares) y la variable factor (por ejemplo, tabaquismo), controlando la tercera variable (por ejemplo, dieta: "alta o baja en grasas"). El estadístico de Cochran es el siguiente:

$$X^2_{\text{Cochran}} = \frac{\left(\sum_k n_k - \sum_k m_k\right)^2}{\sum_k \sigma^2_{n_k}}$$

donde:

k = cada uno de los estratos

n_k = frecuencia observada en cualquier casilla del estrato k (sólo una y siempre la misma en todos los estratos)

m_k = frecuencia esperada correspondiente a n_k

$\sigma^2_{nk} = n_{1+k}\, n_{2+k}\, n_{+1k}\, n_{+2k}/n^3$

($n_{1+k}\, n_{2+k}\, n_{+1k}\, n_{+2k}$ son la cuatro frecuencias marginales asociadas a las tablas 2 x 2 de cada estrato). El estadístico de Mantel-Haenszel es como el de Cochran, pero utiliza, primero la corrección por continuidad (resta 0,5 al numerador antes de elevar al cuadrado) y, segundo, cambia el denominador de la varianza, con $n^2(n-1)$ en vez de n^3. Ambos estadísticos, se distribuyen según χ^2 con 1 grado de libertad. Probabilidades asociadas inferiores a 0,05 llevan a rechazar la hipótesis de independencia condicional una vez controlados la influencia de lo estratos.

9.4 Contenido de las casillas

Hasta ahora se ha visualizado únicamente frecuencias absolutas en las tablas de contingencia, pero se puede visualizar más información, eligiendo para ello en el cuadro de diálogo correspondiente que se muestra en la Figura 9.4.

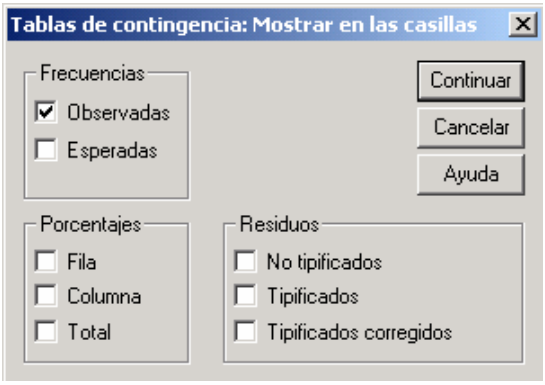


Figura 9.4 Contenido de las casillas en las Tablas de contingencia

Lo que al lector le puede resultar menos familiar es lo referente a los residuos. Respecto de los **no tipificados**, simplemente es la diferencia entre las frecuencias observadas y las esperadas. Los residuos tipificados, es un residuo no tipificado dividido por la raíz cuadrada de su correspondiente frecuencia esperada. Su promedio es 0 pero su desviación típica es inferior a 1 por lo que no sirve para interpretarlos como puntuaciones Z, pero si valen para indicar el grado en que cada casilla contribuye al valor del estadístico *chi-cuadrado*, valor que se obtiene sumando todos los residuos tipificados. Por último, los residuos **tipificados corregidos**, se distribuyen normalmente con media 0 y desviación típica 1, y se obtienen dividiendo el residuo de cada casilla por su error típico.

10. Contraste de hipótesis para una y dos muestras

10.1 Introducción

El objetivo del análisis estadístico es tomar decisiones sobre el conjunto de la población sirviéndose de los resultados de las muestras que se extraen de esa población. En los capítulos precedentes, fundamentalmente se ha procedido a la descripción de las muestras, aunque también hemos realizado alguna incursión en el terreno de la inferencia cuando se determinaba si una variable era o no normal o la variabilidad de la variable dependiente en los diferentes grupos de un factor eran o no homogénea.

En esta tema vamos a estudiar, en primer lugar, un procedimiento descriptivo, denominado Medias, que permite obtener estadísticos descriptivos de los distintos grupos y subgrupos definidos por una o más variables independientes; también veremos el contraste de hipótesis para una muestra, y el contraste de hipótesis para dos muestras, tanto independientes como relacionadas, mediante las denominadas prueba T. Los diversos contrastes de hipótesis para más de dos muestras, lo que se conoce como Análisis de varianza, los veremos en los dos siguientes capítulos.

Los contrastes de hipótesis para una y dos muestras basados en la prueba T tienen la misma estructura, en el sentido de que el estadístico empleado es una tipificación, en laual el numerador es la diferencia entre el valor del estadístico de la muestra y el valor del parámetro de la población de la que supuestamente se ha extraído la muestra, y en el denominador está el error típico de la distribución muestral del estadístico que estemos contrastado (de la media o de la diferencia de medias).

10.2 Medias

Como se ha indicado este procedimiento permite obtener estadísticos descriptivos de una variable independiente teniendo en cuenta los grupos definidos por una o más variables independientes. De manera opcional se puede realizar un Análisis de varianza de un factor, obtener el coeficiente de determinación o proporción de varianza explicada y contrastar la hipótesis de linealidad. Para acceder al procedimiento se sigue la secuencia:

Analizar → Comparar medias → Medias...

y se muestra el cuadro de diálogo de la Figura 10.1

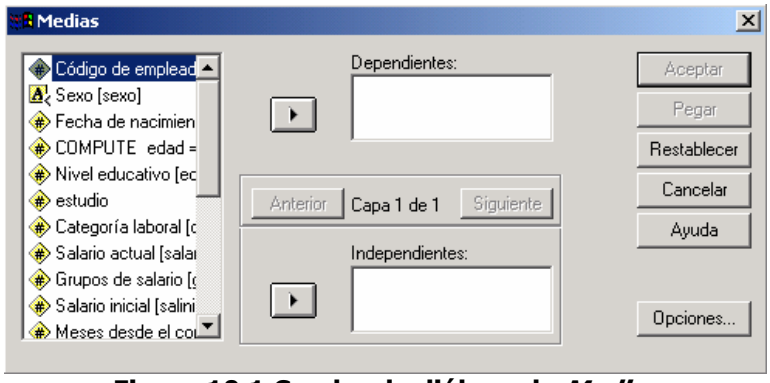


Figura 10.1 Cuadro de diálogo de *Medias*

Para efectuar el análisis se traslada a la lista **Dependientes** la/s variable/s que queramos analizar, y a la lista **independientes** el/los factor/es que actúan como variables independientes. Si se seleccionan como independientes más de una variable, los resultados de los estadísticos de la variable dependiente estarán anidados. Mediante el botón **Opciones**, cuyo cuadro se muestra en la Figura 10.2, se pueden elegir los estadísticos en la lista correspondiente. También se pueden obtener las tablas del Anova y los coeficientes de correlación de Pearson y su cuadrado (proporción de varianza asociada), y el coeficiente de correlación eta y su cuadrado y el contraste de linealidad.

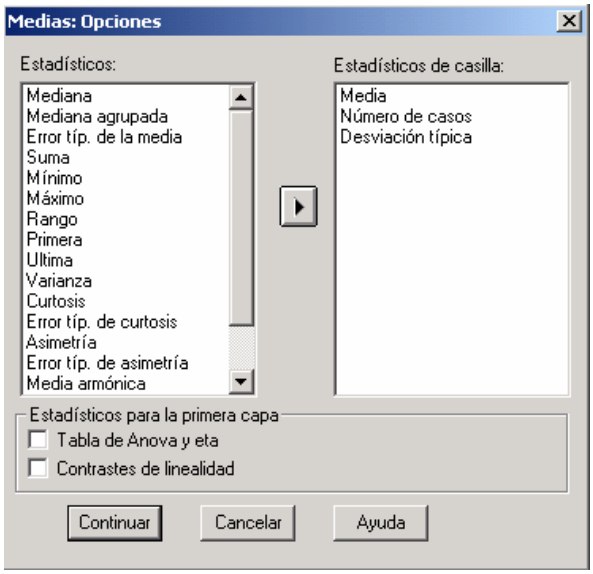


Figura 10.2 Cuadro de diálogo *Opciones* de *Medias*

La Tabla 10.1 muestra los estadísticos requeridos para la variable **educ** (nivel educativo) para cada subgrupo de **catlab** (categoría laboral) y **minoría** (clasificación étnica) del archivo *Datos de empleados*. Por defecto, los estadísticos que se muestran son los que aparecen en la lista **Estadísticos de casilla**, y son: Media, Número de casos y Desviación típica.

Tabla 10.1 Tabla de Estadísticos de *Medias*

Nivel educativo

Clasificación étnica		Categoría laboral			
		Administrativo	Seguridad	Directivo	Total
No	Media	12,82	10,29	17,31	13,69
	N	276	14	80	370
	Desv. típ.	2,36	2,05	1,52	2,94
	Mediana	12,00	12,00	17,00	15,00
	Mediana agrupada	13,16	10,29	16,93	14,09
	Error típ. de la media	,14	,55	,17	,15
	Suma	3538	144	1385	5067
	Mínimo	8	8	15	8
	Máximo	19	12	21	21
	Rango	11	4	6	13
	Primero	8	8	15	8
	Último	19	12	21	21
	Varianza	5,582	4,220	2,319	8,657
	Curtosis	-,153	-2,241	-1,317	-,297
	Error típ. de la curtosis	,292	1,154	,532	,253
	Asimetría	-,510	-,325	,312	-,143
	Error típ. de la asimetría	,147	,597	,269	,127
	Media armónica	12,31	9,88	17,18	12,98
	Media geométrica	12,58	10,09	17,25	13,35
	% de la suma total	55,3%	2,3%	21,7%	79,2%
	% del total de N	58,2%	3,0%	16,9%	78,1%
Sí	Media	13,02	10,08	16,00	12,77
	N	87	13	4	104
	Desv. típ.	2,24	2,47	2,94	2,56
	Mediana	12,00	8,00	16,50	12,00
	Mediana agrupada	13,22	10,00	16,50	12,96
	Error típ. de la media	,24	,68	1,47	,25
	Suma	1133	131	64	1328
	Mínimo	8	8	12	8
	Máximo	18	15	19	19
	Rango	10	7	7	11
	Primero	8	8	12	8
	Último	18	15	19	19
	Varianza	5,023	6,077	8,667	6,529
	Curtosis	,092	-,992	1,500	-,220
	Error típ. de la curtosis	,511	1,191	2,619	,469
	Asimetría	-,353	,606	-,941	-,239
	Error típ. de la asimetría	,258	,616	1,014	,237
	Media armónica	12,59	9,57	15,55	12,20
	Media geométrica	12,82	9,81	15,78	12,49
	% de la suma total	17,7%	2,0%	1,0%	20,8%
	% del total de N	18,4%	2,7%	,8%	21,9%
	% del total de N	76,6%	5,7%	17,7%	100,0%

10.3 Prueba T para una muestra

Esta prueba permite contrastar hipótesis sobre la media poblacional a partir de la media obtenida en la muestra. Cuando se conoce la varianza de la población, el estadístico de contraste es Z, cuya distribución es normal con media 0 y desviación típica 1, pero lo habitual es desconocer la varianza muestral por lo cual es preciso estimarla a partir de la varianza insesgada de la muestra (la cuasivarianza). En estas condiciones, el estadístico de contraste es T, cuya expresión es:

$$T = \frac{\overline{Y} - \mu}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y} - \mu}{S_{n-1}/\sqrt{n}}$$

que se distribuye según el modelo *t* de *Student* con *n-1* grados de libertad. Para que este estadístico se ajuste a este modelo de probabilidad es necesario que la población de la que se ha extraído la muestra sea *normal*, o bien que el tamaño de la muestra sea lo suficientemente grande como para poder obviar el hecho de que la población de referencia no sea normal. Para acceder al procedimiento seguir la secuencia

Analizar → Comparar medias → Prueba T para una muestra...

y se muestra el cuadro de diálogo de la Figura 10.3.

Figura 10.3 Cuadro de diálogo de *Prueba T para una muestra*

Se elige la variable o variable que se desea contrastar y se traslada a la lista **Contrastar variables**, y en **Valor de prueba** se escribe el valor de la media en la población. Para cada variable seleccionada se genera una prueba T y su correspondiente significación (contraste bilateral). Este valor indica la probabilidad de que la muestra contrastada provenga de una población cuya media es el **Valor de Prueba**. Si la probabilidad es muy pequeña (menor de 0,05) se rechaza la hipótesis y viceversa.

La tabla de resultados también ofrece el intervalo de confianza construido sobre la diferencia entre la media muestral (la de la variable) y el Valor de prueba (por defecto el intervalo se construye al 95%). Si este intervalo de confianza contiene el valor 0 no se puede rechazar la hipótesis. Como recordará el lector, estos intervalos se obtiene sumando y restando a la diferencia entre la media muestral y la poblacional, el resultado de multiplicar el error típico de la media ($S_{n-1}/\sqrt{n}$) por el percentil 97,5 de la distribución t para los grados de libertad pertinentes. Para una

tamaño muestral como el del archivo *Datos de empleados*, 474, la distribución para obtener el percentil sería la normal (el valor de este percentil es 1,96).

Si contrastamos para la variable **educ** que el promedio de años de estudio en la población es 13 (Valor de prueba) el resultado obtenido se puede ver en la Tabla 10.2.

Tabla 10.2 Resultados de la Prueba T para una muestra

Estadísticos para una muestra

	N	Media	Desviación típ.	Error típ. de la media
Nivel educativo	474	13,49	2,88	,13

Prueba para una muestra

	Valor de prueba = 13					
					95% Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
Nivel educativo	3,710	473	,000	,49	,23	,75

En la primera tabla se muestra los valores de la variable, y en la segunda el resultado del contraste. El probabilidad del valor de T es menor de 0,05 por lo cual no podemos aceptar la hipótesis que la muestra de 474 sujetos provienen de una población cuyo promedio de años de estudio es 13 (Valor de prueba). A la misma conclusión se llega observando los intervalos de confianza para la diferencia entre la media muestral y la de la población (13,49 – 13 = 0,49). Sobre esta diferencia se ha construido el intervalo de confianza. Siendo el error típico de la media 0,13 (tabla superior de la Tabla 10.2), tendremos,

Límite inferior: 0,49 – 0,13 x 1,96 = 0,23

Límite superior: 0,49 + 0,13 x 1,96 = 0,75.

10.4 Prueba T para dos muestras independientes

Esta prueba permite contrastar hipótesis de que las medias de dos poblaciones independientes (μ_1 y μ_2) son iguales, utilizando para ello las medias, $\overline{Y}_1$ e $\overline{Y}_2$, de dos muestras aleatorias, de tamaño n_1 y n_2, extraídas de esas poblaciones.

El estadístico T que se utiliza para el contraste tiene la siguiente estructura:

$$T = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - \left(\mu_1 - \mu_2\right)}{\hat{\sigma}_{\overline{Y}_1 - \overline{Y}_2}}$$

si se suponen las varianzas poblacionales iguales $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$ el error típico de la diferencia de medias es:

$$\hat{\sigma}_{\overline{Y}_1 - \overline{Y}_2} = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde $\hat{\sigma}$ se estima a través de la raíz cuadrada de la media ponderada de las varianzas insesgadas muestrales, y su expresión es:

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)S^2_{n_1-1} + (n_2 - 1)S^2_{n_2-1}}{n_1 + n_2 - 2}}$$

distribuyéndose T según el modelo *t* de *Student* con n_1+n_2-2 grados de libertad.

Si no se pueden suponer las varianzas poblacionales iguales, entonces cada una de las varianzas poblacionales hay que estimarlas mediante las varianzas insesgadas muestrales, y el error típico será:

$$\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\frac{S^2_{n_1-1}}{n_1} + \frac{S^2_{n_2-1}}{n_2}}$$

y aunque en estas condiciones T se distribuye según el modelo *t* de *Student*, los grados de libertad de la distribución necesitan ser estimados mediante la ecuación propuesta por Welch (1938):

$$gl = \frac{\left(\frac{S^2_{n_1-1}}{n_1} + \frac{S^2_{n_2-1}}{n_2}\right)^2}{\frac{\left(\frac{S^2_{n_1-1}}{n_1}\right)^2}{n_1-1} - \frac{\left(\frac{S^2_{n_2-1}}{n_2}\right)^2}{n_2-1}}$$

Para tomar una decisión la igualdad de varianzas de las poblaciones, este procedimiento muestra las dos versiones del estadístico T, y también la prueba de *Levene* sobre igualdad de varianzas, sobre cuyo resultado se tomará la decisión acerca de la igualdad.

Para acceder a este contraste se sigue la secuencia:

Analizar → Comparar medias → Prueba T para dos muestras independientes

y se muestra el cuadro de diálogo de la Figura 10.4.

Figura 10.4. Cuadro de diálogo de Prueba T para muestras independientes

A la lista **Contrastar variables** se pasan todas las variables dependientes que se deseen contrastar, y al cuadro **Variable de agrupación** se traslada la variable que define los dos grupos. Esta variable puede tener formato numérico o de cadena corta. Una vez elegida, se tienen que **Definir los grupos**, pulsando el botón correspondiente. En el cuadro de diálogo que se muestra en la figura 10.5, se definen o bien los **valores de los grupos**, o bien, si la variable es cuantitativa, se puede especificar el **punto de corte**: los casos con puntuación mayor o igual que dicho punto de corte forman un grupo y los de menor valor forman otro grupo. Esta opción sólo se muestra cuando la variable elegida para formar los grupos es de tipo numérico.

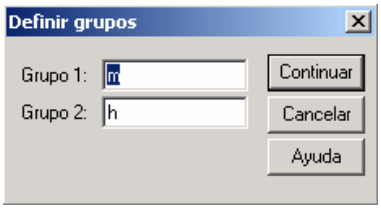


Figura 10.5 Cuadro para Definir grupos

Para ilustrar el resultado se ha contrastado la hipótesis de que, en la población, el promedio de años de estudio (variable **educ**) de hombres y mujeres (variable **sexo**) es el mismo. Las tablas que se generan son las que se muestran en la Tabla 10.3

Tabla 10.3 Resultados de la Prueba T para muestras independientes

Estadísticos de grupo

	Sexo	N	Media	Desviación típ.	Error típ. de la media
Nivel educativo	Mujer	216	12,37	2,32	,16
	Hombre	258	14,43	2,98	,19

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Nivel educativo	Se han asumido varianzas iguales	17,884	,000	-8,276	472	,000	-2,06	,25	-2,55	-1,57
	No se han asumido varianzas iguales			-8,458	469,595	,000	-2,06	,24	-2,54	-1,58

Se observa que la hipótesis de igualdad de varianzas no es posible aceptarla (significación menor de 0,05 del estadístico de *Levene*), por lo cual tenemos que ver el valor de T calculado sin asumir ese supuesto, y cuyo resultado nos indica que los promedios de años de estudio no son iguales en la población de hombres y mujeres, pues la significación del valor de T obtenido es inferior a 0,05, y por tanto el intervalo de confianza no contiene el valor cero.

10.5 Prueba T para dos muestras relacionadas

Esta prueba permite contrastar hipótesis sobre igualdad entre dos medias relacionadas. Es decir se tiene una población de diferencias con media μ_D, resultado de restar las puntuaciones de un mismo grupo en dos variables diferentes o en la misma variable en dos momentos diferentes. De la población de diferencias se extrae un muestra aleatoria de tamaño *n* y se utiliza la media de esa muestra, $\overline{Y}_D$, para contrastar la hipótesis de que la media de la población de diferencias vale 0. El estadístico T, sigue teniendo la misma estructura y su expresión es:

$$T = \frac{\overline{Y}_D - \mu_D}{\hat{\sigma}_{\overline{Y}_D}} = \frac{\overline{Y}_D - \mu_D}{S_D}$$

siendo S_D la desviación típica insesgada de la *n* diferencias, e igual a $\sqrt{\frac{S_1^2 + S_2^2 - 2S_{12}}{n}}$ El estadístico se distribuye según el modelo *t* de *Student*, con *n – 1* grados de libertad.

Para que el valor T se ajuste a este modelo de forma apropiada es necesario que la población de diferencias se distribuya normalmente.

Se accede al procedimiento mediante

Analizar → Comparar medias → Prueba T para dos muestras relacionadas

y se muestra el cuadro de la Figura 10.6

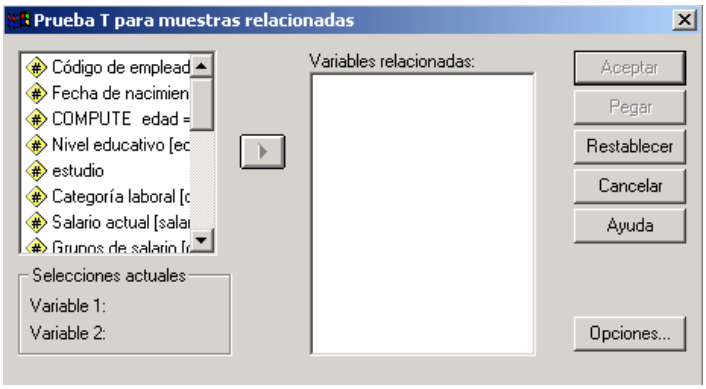


Figura 10.6 Cuadro de diálogo de Prueba T para dos muestras relacionadas

La lista de variables sólo incorpora las que tienen formato numérico. Para ejecutar el procedimiento se trasladan a la lista **Variables relacionadas** las variables por parejas que se desean contrastar (hasta que no se han marcado dos variables no se activa la flecha de paso de una lista a otra). Se pueden elegir tantos pares de variables como se deseen.

Para ilustrar el resultado, se ha contrastado la hipótesis (siendo conscientes del resultado que se va a obtener) de que el salario inicial es igual al salario actual. En la Tabla 10.4 se muestran las tablas que se generan en este procedimiento.

Tabla 10.4 Resultados de la Prueba T para muestras relacionadas

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	Salario actual	$34,419.57	474	$17,075.66	$784.31
	Salario inicial	$17,016.09	474	$7,870.64	$361.51

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	Salario actual y Salario inicial	474	,880	,000

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia			
					Inferior	Superior		
Par 1	Salario actual - Salario inicial	$17,403.48	$10,814.62	$496.73	$16,427.41	$18,379.56	35,036	473

La primera tabla recoge los valores de Media, Número de casos, Desviación típica y el Error típico de la media. La segunda informa del coeficiente de correlación de

Pearson y su significación estadística. Por último, la tercera tabla es la del contraste propiamente dicho e indica el valor del estadístico T y su significación. En este caso, el resultado (ya previsto) es que la diferencia de salarios es significativamente distinta de cero.

11. Análisis de varianza de un factor

11.1 Introducción

El Análisis de varianza (ANOVA) permite comparar varios grupos de una variable cuantitativa. A la variable categórica u ordinal que define los grupos se le denomina variable independiente (VI) o factor y a la variable cuantitativa se le denomina variable dependiente (VD) o variable de respuesta.

Para realizar un ANOVA en SPSS deberemos tener al menos una variable independiente o de agrupamiento, con más de dos categorías u ordenes, y una variable dependiente. Mediante la técnica del ANOVA contrastamos la hipótesis nula de que los promedios de la VD respecto de un factor o VI con más de dos grupos o niveles son iguales, frente a la hipótesis alternativa de que al menos el promedio en un grupo es diferente a los demás.

11.2 ANOVA de un factor

Para llevar a cabo el ANOVA se utiliza el estadístico F, que es un cociente que refleja la relación que hay entre la variabilidad en la población de los promedios entre los grupos (numerador) y la variabilidad en la población dentro de los grupos (denominador). Si los promedios de los grupos en la población son iguales, las medias muestrales serán similares y las diferencias que se puedan detectar serán atribuibles al azar. En este caso la estimación del numerador de F reflejará una variabilidad similar a la que refleje el denominador, variabilidad ésta basada en las diferencias individuales. El resultado de F será pues próximo a 1. Si, por el contrario, las medias muestrales son distintas, la estimación de su variabilidad tendrá un mayor grado de variabilidad que el de las propias diferencias individuales, y el resultado del cociente será entonces mayor que 1, y cuanto más diferentes sean las medias de los grupos mayor será el valor de F (una explicación detallada del fundamento del ANOVA se puede encontrar en Keppel, 1991).

Al igual que ocurre con la prueba T, el valor del estadístico F será un percentil dentro de la distribución *F* de *Fisher-Snedecor* con grados de libertad los del numerador (número de grupos menos 1) y los del denominador (número total de sujetos menos número de grupos) , y por tanto tendrá una probabilidad asociada, la cual indicará si se acepta la hipótesis nula de igualdad de medias o se rechaza, en caso de que la probabilidad del valor de F sea inferior a 0,05. Para que el estadístico F se distribuya según el modelo F de *Fisher-Snedecor*, es preciso que se cumplan dos supuestos básicos: 1) que las poblaciones de las que se han obtenido las muestras sean normales, y 2) que las varianzas sea iguales (supuesto de homocedasticidad).

Si se rechaza la hipótesis nula de igualdad de medias, implica que al menos una de las medias es diferente al resto. Para determinar entre qué grupos de medias se dan las diferencias es preciso proceder a comparaciones entre las medias. Estas comparaciones pueden ser de dos tipos: comparaciones planificada o *a priori*, o comparaciones *post hoc*. Las primeras permiten comparaciones más complejas y focalizadas que las segundas, en el sentido de que éstas sólo comparan las

diferencias dos a dos entre las medias de todos los grupos, mientras que aquéllas permiten, por ejemplo, comparar si la media de un grupo es igual al promedio de las medias de los restantes grupos.

11.3 El procedimiento ANOVA de un factor

Para ilustrar el proceso del procedimiento vamos a basarnos en el siguiente conjunto de datos que representan las puntuaciones obtenidas por 15 escolares en una prueba de comprensión lectora (VD), bajo tres diferentes tipos de instrucción (grupos o niveles del Factor o VI). Al primer grupo se le pide que memorice un ensayo, al segundo se le pide que se concentre en las ideas que contiene el ensayo, y al tercero no se le da ninguna instrucción específica. La puntuación obtenida por cada sujeto es el número de ítems de un test respondidos correctamente. Cada grupo está compuesto por 5 sujetos, asignados de manera aleatoria. Las puntuaciones se pueden en la siguiente tabla:

Factor A		
Nivel 1	**Nivel 2**	**Nivel 3**
16	**4**	**2**
18	**6**	**10**
10	**8**	**9**
12	**10**	**13**
19	**2**	**11**

Estos datos en el editor de datos de SPSS estarían reflejados en dos únicas variable: la VI o Factor, con tres valores (1, 2 y 3) y la variable de respuesta. Tal como se muestra en la Figura 11.1

	instruc	aciertos
1	1	16
2	1	18
3	1	10
4	1	12
5	1	19
6	2	4
7	2	6
8	2	8
9	2	10
10	2	2
11	3	2
12	3	10
13	3	9
14	3	13
15	3	11

Figura 11.1 Datos de una VI y una VD para el ANOVA de un factor con muestras independientes

Para acceder al procedimiento ANOVA se sigue la secuencia

Analizar → Comparar medias → ANOVA de un factor...

y se muestra el cuadro de la Figura 11.2

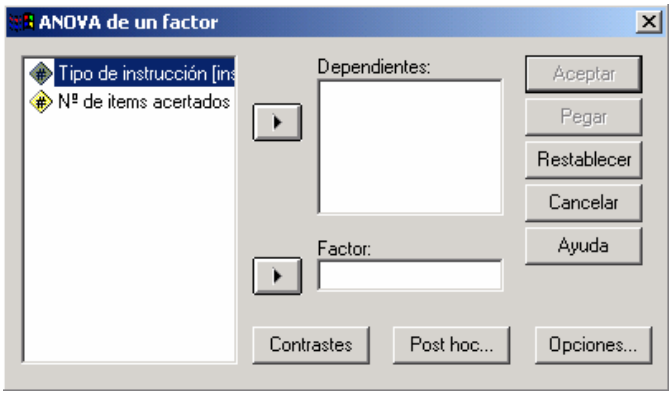


Figura 11.2 Cuadro de diálogo ANOVA de un factor

La lista de variables muestra todas las variables del archivo con formato numérico (están excluidas las variables con formato de cadena). A la lista **Dependientes** se trasladan todas las VD's que queremos analizar y al cuadro **Factor** se traslada la VI que es la que define los grupos. Sin realizar más especificaciones que las que tiene por defecto el programa, la tabla resumen del ANOVA es la que se muestra en la Tabla 11.1

Tabla 11.1 Tabla resumen del procedimiento *ANOVA de un factor*

ANOVA

Nº de items acertados

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	210,000	2	105,000	7,412	,008
Intra-grupos	170,000	12	14,167		
Total	380,000	14			

El estadístico F es el cociente de las dos medias cuadráticas, la Inter-grupos y la Intra-grupos, Ambas medias cuadráticas son dos estimadores diferentes e independientes de la varianza poblacional. La probabilidad asociada al valor es menor de 0,05, por lo que se rechaza la hipótesis de que la poblaciones definidas por la variable "Tipo de instrucción" no tiene la misma media respecto de la VD.

Esta tabla resumen es la opción mínima del procedimiento ANOVA, pero hay una serie de ellas más a las que se accede pulsando el correspondiente botón del cuadro de diálogo, y que muestra el cuadro de la Figura 11.3

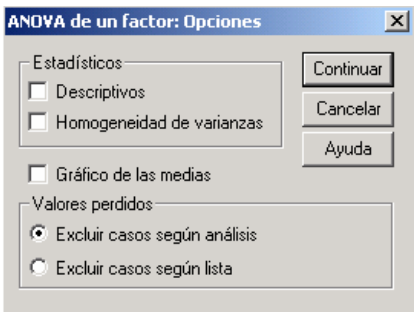


Figura 11.3 Cuadro de opciones de ANOVA de un factor

En el recuadro Estadísticos se incluyen algunos estadísticos descriptivos y la prueba de *Levene* de homogeneidad de varianzas, ya comentada en el capítulo anterior.

También se puede obtener un gráfico de líneas, con la variable factor en el eje de abcisas y la variable dependiente en el de ordenadas. Además de este estadístico, SPSS dispone del gráfico denominado **Barras de error**, que es muy útil para visualizar los datos de un ANOVA de un factor. Al final de ese capítulo puede verse este gráfico comentado sobre los datos que nos están sirviendo para ilustrar el procedimiento.

Las tablas que proporcionan las opciones del recuadro Estadísticos son las que se muestran en la Tabla 11.2.

Tabla 11.2 Tablas resumen de Estadísticos y prueba de homocedasticidad

Descriptivos

Nº de items acertados

		Memorizar ensayo	Fijarse en ideas ensayo	Sin instrucción	Total
N		5	5	5	15
Media		15,00	6,00	9,00	10,00
Desviación típica		3,87	3,16	4,18	5,21
Error típico		1,73	1,41	1,87	1,35
Intervalo de confianza para la media al 95%	Límite inferior	10,19	2,07	3,81	7,11
	Límite superior	19,81	9,93	14,19	12,89
Mínimo		10	2	2	2
Máximo		19	10	13	19

Prueba de homogeneidad de varianzas

Nº de items acertados

Estadístico de Levene	gl1	gl2	Sig.
,189	2	12	,830

Por la significación del valor de la prueba de Levene, se puede afirmar que las varianzas son homogéneas.

Por último, el gráfico de medias para estos datos es el que se muestra en la Figura 11.4.

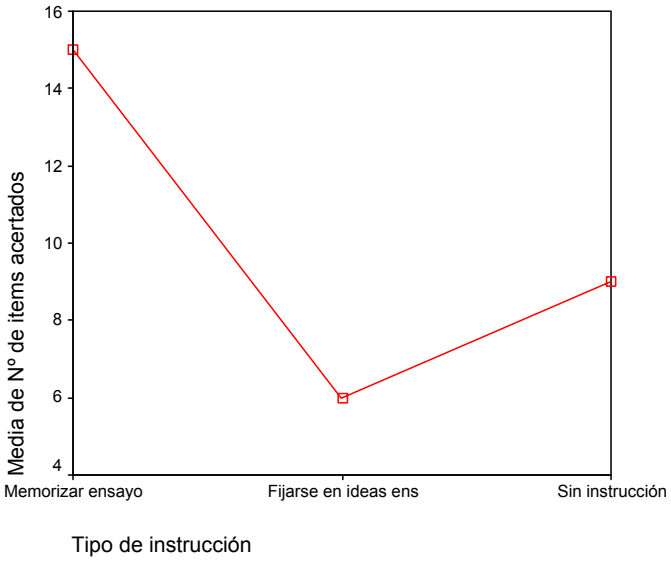


Figura 11.4 Gráfico de medias del procedimiento ANOVA

11.5 Comparaciones múltiples *a posteriori* o *post hoc*

Cuando el resultado del ANOVA resulta significativo, es preciso detectar entre qué medias poblacionales se dan las diferencias. Con los contrastes *a posteriori* se comparan entre sí, dos a dos, todas las medias de los grupos del factor. Para elegir el tipo de contraste se pulsa en el botón **Post hoc...** del cuadro del ANOVA y se muestra el cuadro de la Figura 11.5.

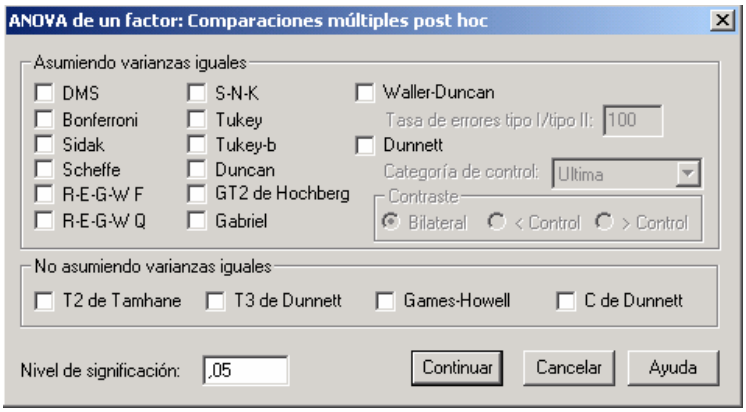


Figura 11.5 Cuadro de comparaciones múltiples de ANOVA

Se puede elegir entre los tipos de comparaciones que se muestran en el cuadro según las varianzas sean o no iguales. Las diferencias entre unos métodos y otros estriban fundamentalmente en la distribución de probabilidad en la que se basan, en cómo controlan la tasa de error de las comparaciones que efectúan, y en el procedimiento como se llevan a cabo las comparaciones.

Para todos estos métodos de comparación múltiple, se puede establecer el Nivel de significación, que por defecto tiene el valor de 0,05. El lector puede ver en la Tabla 11.3, el resultado de las comparaciones múltiples de todas las opciones disponibles asumiendo varianzas iguales.

Tabla 11.3. Todos los tipos de comparaciones múltiples asumiendo varianzas iguales del procedimiento ANOVA

Comparaciones múltiples

Variable dependiente: Nº de items acertados

	(I) Tipo de instrucción	(J) Tipo de instrucción	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
HSD de Tukey	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,007	2,65	15,35
		Sin instrucción	6,00	2,38	,065	-,35	12,35
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,007	-15,35	-2,65
		Sin instrucción	-3,00	2,38	,443	-9,35	3,35
	Sin instrucción	Memorizar ensayo	-6,00	2,38	,065	-12,35	,35
		Fijarse en ideas ensayo	3,00	2,38	,443	-3,35	9,35
Scheffé	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,009	2,36	15,64
		Sin instrucción	6,00	2,38	,078	-,64	12,64
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,009	-15,64	-2,36
		Sin instrucción	-3,00	2,38	,474	-9,64	3,64
	Sin instrucción	Memorizar ensayo	-6,00	2,38	,078	-12,64	,64
		Fijarse en ideas ensayo	3,00	2,38	,474	-3,64	9,64
DMS	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,003	3,81	14,19
		Sin instrucción	6,00*	2,38	,027	,81	11,19
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,003	-14,19	-3,81
		Sin instrucción	-3,00	2,38	,232	-8,19	2,19
	Sin instrucción	Memorizar ensayo	-6,00*	2,38	,027	-11,19	-,81
		Fijarse en ideas ensayo	3,00	2,38	,232	-2,19	8,19
Bonferroni	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,008	2,38	15,62
		Sin instrucción	6,00	2,38	,081	-,62	12,62
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,008	-15,62	-2,38
		Sin instrucción	-3,00	2,38	,695	-9,62	3,62
	Sin instrucción	Memorizar ensayo	-6,00	2,38	,081	-12,62	,62
		Fijarse en ideas ensayo	3,00	2,38	,695	-3,62	9,62
Sidak	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,008	2,41	15,59
		Sin instrucción	6,00	2,38	,079	-,59	12,59
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,008	-15,59	-2,41
		Sin instrucción	-3,00	2,38	,546	-9,59	3,59
	Sin instrucción	Memorizar ensayo	-6,00	2,38	,079	-12,59	,59
		Fijarse en ideas ensayo	3,00	2,38	,546	-3,59	9,59
Gabriel	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,008	2,46	15,54
		Sin instrucción	6,00	2,38	,075	-,54	12,54
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,008	-15,54	-2,46
		Sin instrucción	-3,00	2,38	,526	-9,54	3,54
	Sin instrucción	Memorizar ensayo	-6,00	2,38	,075	-12,54	,54
		Fijarse en ideas ensayo	3,00	2,38	,526	-3,54	9,54
Hochberg	Memorizar ensayo	Fijarse en ideas ensayo	9,00*	2,38	,008	2,46	15,54
		Sin instrucción	6,00	2,38	,075	-,54	12,54
	Fijarse en ideas ensayo	Memorizar ensayo	-9,00*	2,38	,008	-15,54	-2,46
		Sin instrucción	-3,00	2,38	,526	-9,54	3,54
	Sin instrucción	Memorizar ensayo	-6,00	2,38	,075	-12,54	,54
		Fijarse en ideas ensayo	3,00	2,38	,526	-3,54	9,54
t de Dunnett (bilateral)[a]	Memorizar ensayo	Sin instrucción	6,00*	2,38	,048	4,29E-02	11,96
	Fijarse en ideas ensayo	Sin instrucción	-3,00	2,38	,375	-8,96	2,96

*. La diferencia entre las medias es significativa al nivel .05.

a. Las pruebas t de Dunnett tratan un grupo como control y lo comparan con todos los demás grupos.

Se puede ver que el número de comparaciones significativas es diferente según el método que se use, ya que los procedimientos de comparación difieren de un método a otro.

Junto a la tabla con las comparaciones, en el visor también se ofrece una clasificación de los grupos según su parecido entre las medias. Esta tabla se muestra en la Tabla 11.4, y en ella se observan diferencias entre los métodos según los grupos que componen los subconjuntos.

Tabla 11.4 Tablas de subgrupos homogéneos del procedimiento ANOVA

Nº de items acertados

	Tipo de instrucción	N	Subconjunto para alfa = .05	
			1	2
Student-Newman-Keuls [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	
	Memorizar ensayo	5		15,00
	Sig.		,232	1,000
HSD de Tukey [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	9,00
	Memorizar ensayo	5		15,00
	Sig.		,443	,065
Tukey B [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	
	Memorizar ensayo	5		15,00
Duncan [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	
	Memorizar ensayo	5		15,00
	Sig.		,232	1,000
Scheffé [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	9,00
	Memorizar ensayo	5		15,00
	Sig.		,474	,078
Gabriel [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	9,00
	Memorizar ensayo	5		15,00
	Sig.		,526	,075
F de Ryan-Einot-Gabriel-Welsch	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	
	Memorizar ensayo	5		15,00
	Sig.		,232	1,000
Rango de Ryan-Einot-Gabriel-Welsch	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	
	Memorizar ensayo	5		15,00
	Sig.		,232	1,000
Hochberg [a]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	9,00
	Memorizar ensayo	5		15,00
	Sig.		,526	,075
Waller-Duncan [a,b]	Fijarse en ideas ensayo	5	6,00	
	Sin instrucción	5	9,00	
	Memorizar ensayo	5		15,00

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

a. Usa el tamaño muestral de la media armónica = 5,000.

b. Razón de seriedad del error de tipo 1/tipo 2 = 100

El lector puede encontrar en Kirk (1990) una relación exhaustiva de los procedimientos de comparación múltiple, ente los que se encuentran algunos de los que incluye SPSS.

11.5 Comparaciones *planeadas* o *a priori*

En muchas ocasiones, las comparaciones dos a dos entre grupos de un factor no siempre son del interés de los investigadores y necesitan contrastes más complejos. Este tipo de contrastes han de ser planificados y especificados utilizando el botón **Contrastes** del cuadro de diálogo del ANOVA de un factor. El cuadro al que se accede se muestra en la Figura 11.6.

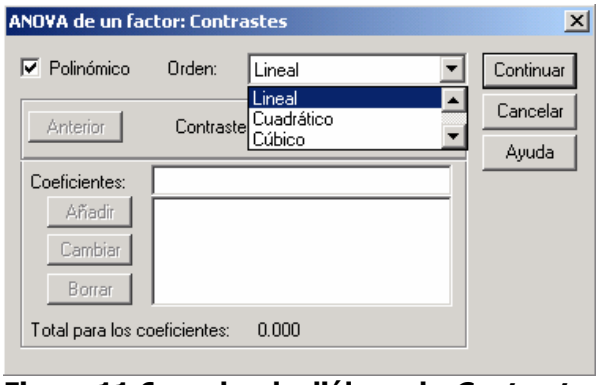


Figura 11.6 cuadro de diálogo de *Contrastes* de *ANOVA de un factor*

- **Polinómicos**. Esta primera opción permite hacer comparaciones de las tendencias. La probabilidad asociada al valor del estadístico F obtenido informará de la aceptación o rechazo de la hipótesis de igualdad de medias. Si la conclusión es de rechazo indicará que hay relación entre la VI y la VD. En el caso de que la VI sea cuantitativa, esta opción permite determinar cuál es el grado de la relación (el máximo calculable es de 5º grado, y se marca en el cuadro **Orden**) entre las dos variables.

  La opción Polinomio ofrece dos soluciones: la no ponderada cuando los niveles del factor están igualmente espaciados y los grupos son equilibrados (mismo tamaño); y la ponderada, cuando los grupos no están igualmente espaciados y/o los grupos no son equilibrados. El número máximo de polinomios que se pueden obtener será igual a los grados de libertad de la suma de cuadrados intergrupos (número de grupos o niveles del factor menos 1), y cada solución polinómica es un componente ortogonal (independiente) de dicha suma de cuadrados.

- **Coeficientes**. Con la opción anterior se contrasta la tendencia de todos los grupos tomados conjuntamente, pero en ocasiones interesa personalizar los contrastes. Para ello se estipulan los coeficientes para determinar los grupos que se desea comparar. Por ejemplo, puede ser útil saber si, para los datos que estamos analizando, la media del grupo que no recibe instrucción es igual al promedio de los otros dos grupos, de modo que los coeficientes asignados podrán ser –1/2, -1/2, 1, o, de forma equivalente, -0,5, -0,5, 1. Si, por ejemplo, quisiéramos determinar si la

media del primer grupo es igual a la suma de las medias de los grupos 2º y 3º, los coeficientes podrían ser 2, -1, -1.

El orden en que se asignan los coeficientes se corresponde con el código ascendente de los grupos del factor o VI, y es preciso asignar tantos coeficientes como grupos, de manera que si se desea que un grupo no intervenga en el contraste se le asigna el valor 0. Cuando el contraste que queremos realizar es de tipo lineal, la suma de los coeficiente de ser 0, y para que dos contraste lineales sean independientes entre sí (ortogonales), es decir, para que dos contrastes no aportan información redundante la suma de los productos de los coeficientes debe valer también cero $\left(\sum c_i c_j = 0\right)$. El número máximo de contraste lineales ortogonales será igual al número de grupos del factor menos uno. Para nuestros datos dos grupos de contrastes ortogonales podrían ser:

-0,5	**-0,5**	**1**
1	**-1**	**0**

con lo que contrastaríamos, por un lado, que el promedio de las medias de los grupos 1 y 2 es igual a la media del grupo 3, y por otro, si la media del grupo 1 es igual a la del grupo 2. Ambos contrastes ofrecen información no redundante, dado su carácter ortogonal (vea el lector que la suma de los productos de los coeficientes vale 0).

Se pueden definir hasta 10 contrastes diferentes con un máximo de 50 coeficientes por contraste. Para definir un nuevo contraste se pulsa en el botón **Siguiente**.

En la Tabla 11.5 se muestra la Tabla del ANOVA para un contraste Polinómico de los datos que no están sirviendo para ilustrar el tema. Recuerde el lector que la VI que estamos utilizando es nominal, por lo que no tiene sentido este contraste, ya que depende del orden en que hemos asignado valores a las etiquetas de la variable. Si el lector varía este orden y realiza el contraste polinómico el resultado sería diferente. Con variables nominales, pues, no tiene sentido este tipo de contrastes. Además, no ofrece soluciones ponderadas porque los valores de los grupos son 1, 2 y 3 y además hay el mismo número de sujetos por grupo.

Tabla 11.5 Tabla resumen de contrastes de tendencias de ANOVA de un factor

ANOVA

Nº de items acertados

			Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	(Combinados)		210,000	2	105,000	7,412	,008
	Término lineal	Contraste	90,000	1	90,000	6,353	,027
		Desviación	120,000	1	120,000	8,471	,013
	Término cuadrático	Contraste	120,000	1	120,000	8,471	,013
Intra-grupos			170,000	12	14,167		
Total			380,000	14			

Como sólo hay tres grupos, el contraste máximo es el cuadrático. Debajo del primer contraste, el del término lineal, aparece la información referente a los contrastes de orden superior no efectuados (Desviación), y el nivel crítico o significación de dichos contrastes. Se observa, como ya se vio en el gráfico de medias de la Figura 11.4 que el término cuadrático también es significativo, es decir hay una tendencia parabólica en los promedios de los grupos (forma de U).

Respecto de los contrastes personalizados hemos realizado los dos ortogonales planteados anteriormente, y el resultado se puede ver en la Tabla 11.6.

Tabla 11.6 Tabla de contrastes de ANOVA de un factor

Coeficientes de los contrastes

	Tipo de instrucción		
Contraste	Memorizar ensayo	Fijarse en ideas ensayo	Sin instrucción
1	-,5	-,5	1
2	1	-1	0

Pruebas para los contrastes

		Contraste	Valor del contraste	Error típico	t	gl	Sig. (bilateral)
Nº de items acertados	Asumiendo igualdad de varianzas	1	-1,50	2,06	-,728	12	,481
		2	9,00	2,38	3,781	12	,003
	No asumiendo igualdad de varianzas	1	-1,50	2,18	-,688	6,909	,514
		2	9,00	2,24	4,025	7,692	,004

La tabla de coeficientes muestra los que se han asignado a cada contraste establecido, y en la tabla de las pruebas, el valor del estadístico T de contraste y su valor crítico, en sus dos modalidades: asumiendo o no la igualdad de varianzas. A la vista de esta tabla no se puede rechazar la hipótesis planteada en el primer contraste (el promedio de las medias de los grupos 1 y 2 es igual a la media del grupo 3) y sí se puede rechazar la hipótesis planteada en el segundo contraste (la media del grupo 1 es diferente a la del grupo 2), pues el nivel crítico es inferior a 0,05.