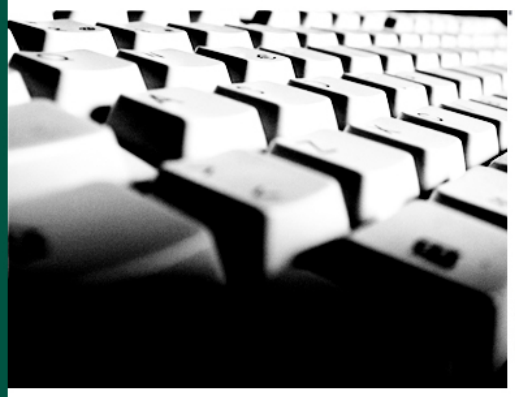
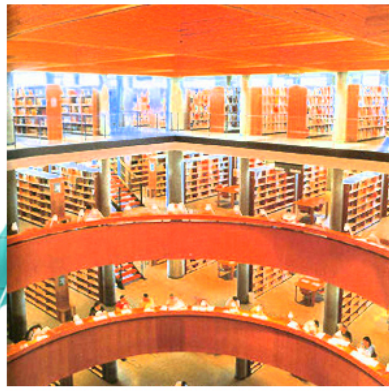
Guías prácticas del IUED

Herramientas de Investigación





Manual de Uso de SPSS

Enrique Moreno González

UNED IUED Instituto Universitario de Educación a Distancia

2008

1ª Edición: Madrid, octubre de 2008

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.

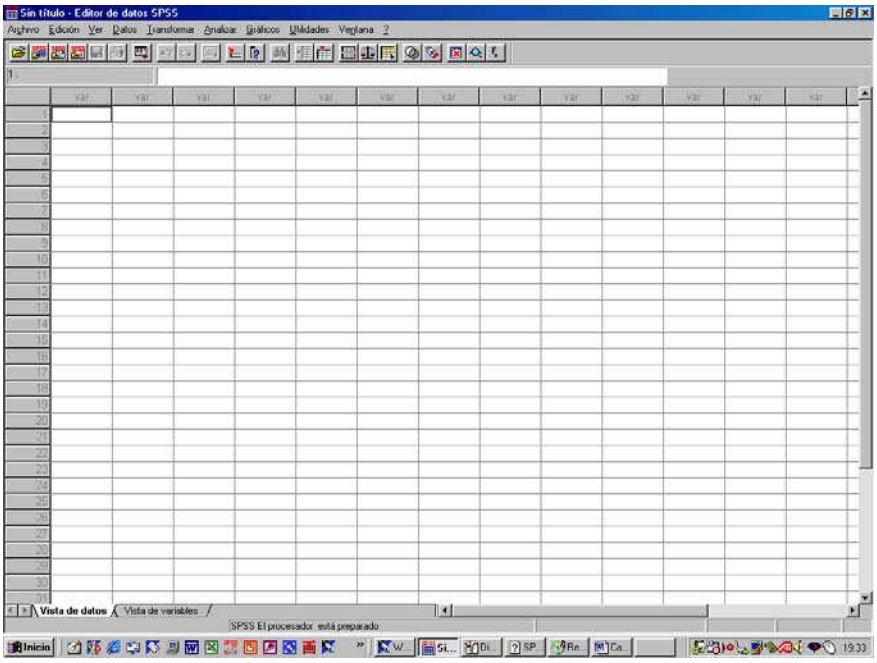


Fig. 1.1 Editor de datos de SPSS, sin datos

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.

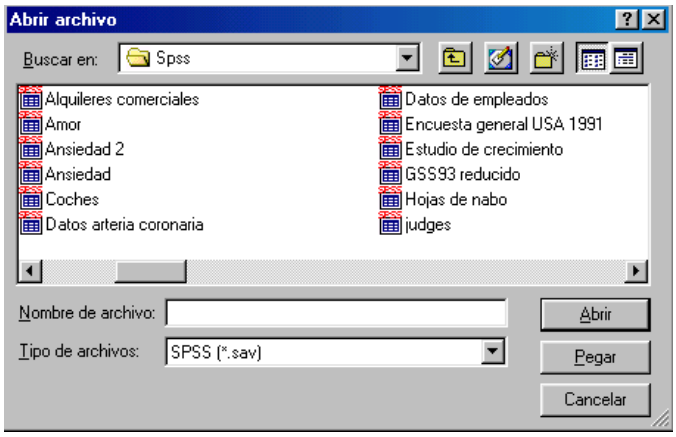
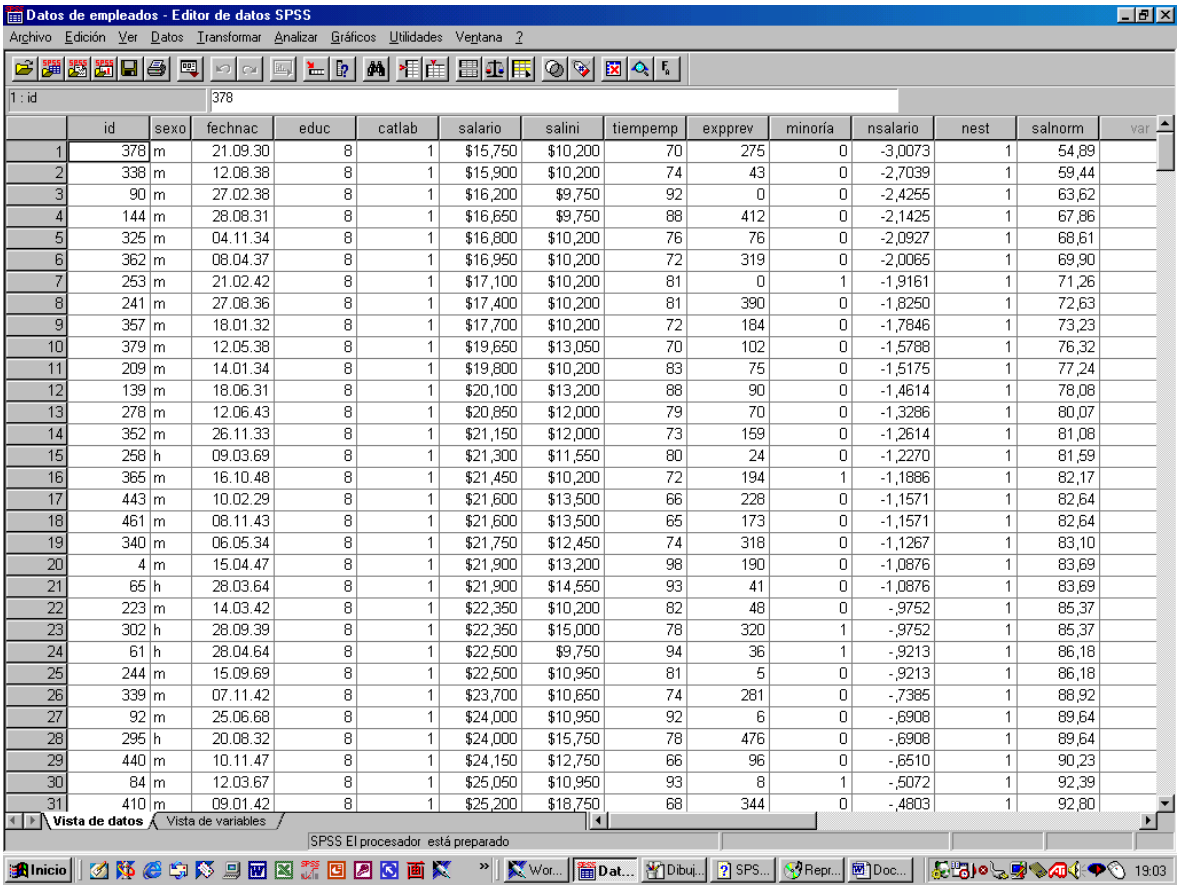


Figura 1.2 Cuadro de diálogo de *Abrir archivo*

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.

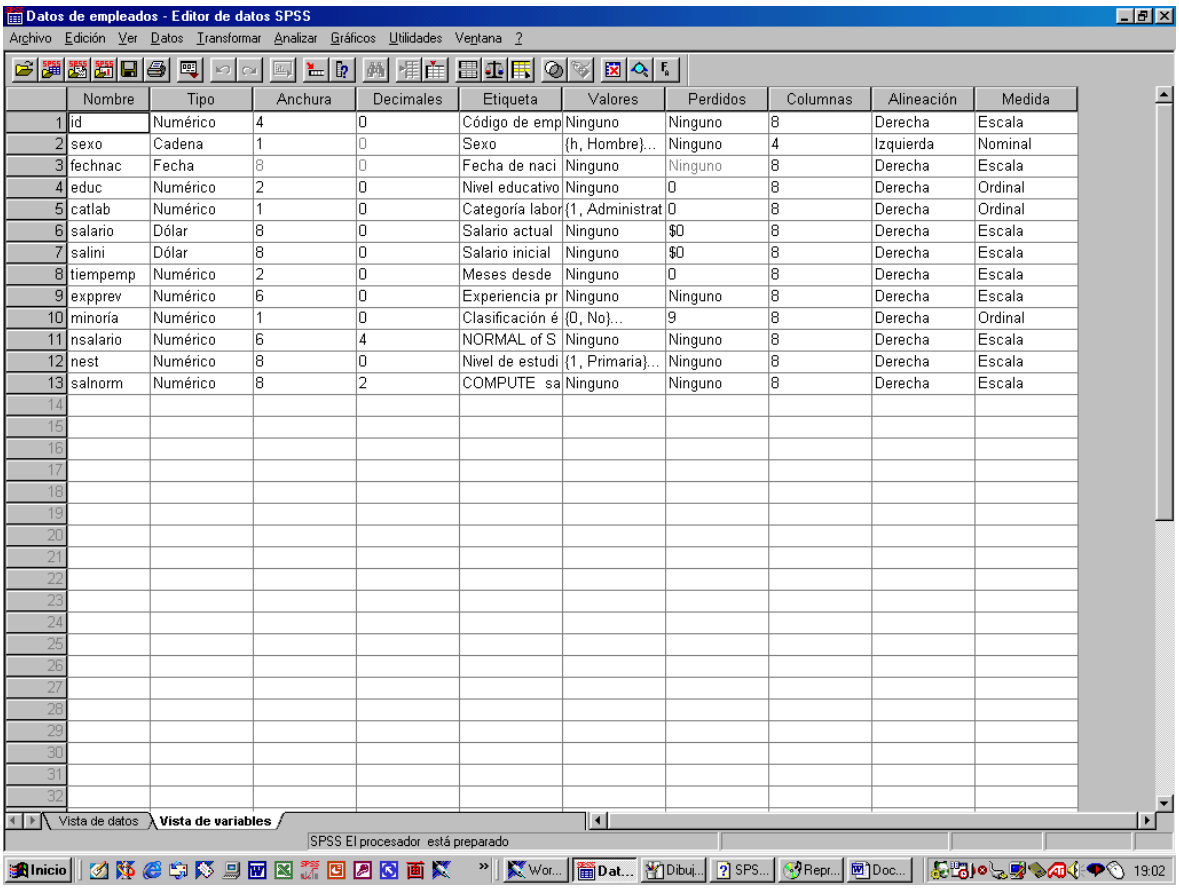


	id	sexo	fechnac	educ	catlab	salario	salini	tiempemp	expprev	minoría	nsalario	nest	salnorm
1	378	m	21.09.30	8	1	$15,750	$10,200	70	275	0	-3,0073	1	54,89
2	338	m	12.08.38	8	1	$15,900	$10,200	74	43	0	-2,7039	1	59,44
3	90	m	27.02.38	8	1	$16,200	$9,750	92	0	0	-2,4255	1	63,62
4	144	m	28.08.31	8	1	$16,650	$9,750	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.

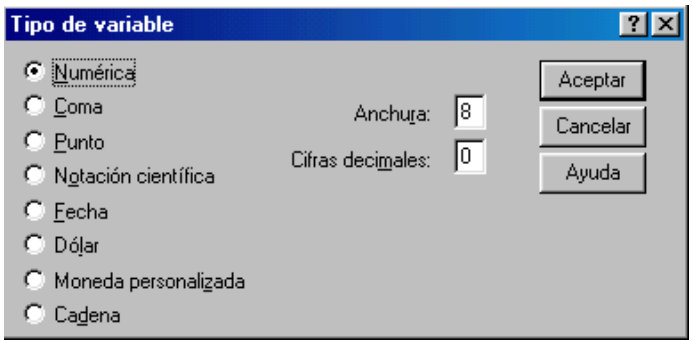


Figura 1.5 Cuadro de definición del tipo de variable

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.

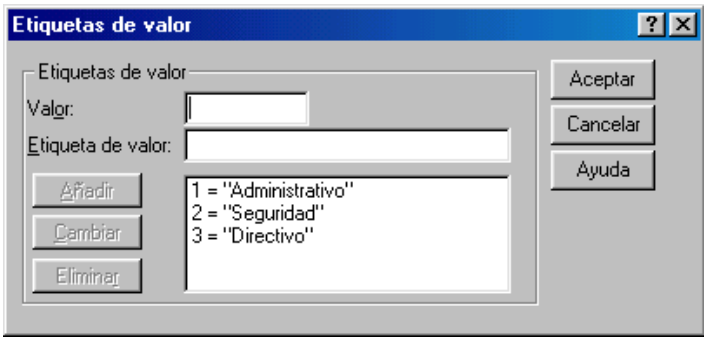


Figura 1.6. Cuadro para etiquetar los valores de variables nominales u ordinales

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.

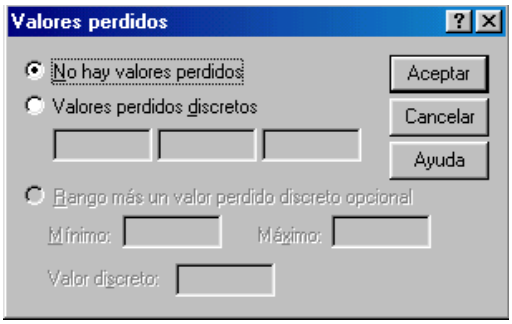


Figura 1.7 Cuadro para definir los valores perdidos

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.

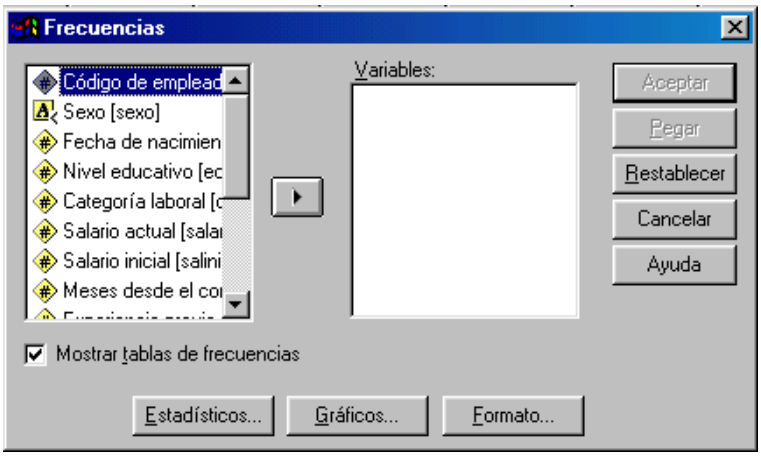


Fig. 1.8 Cuadro de diálogo del procedimiento Frecuencias

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.*

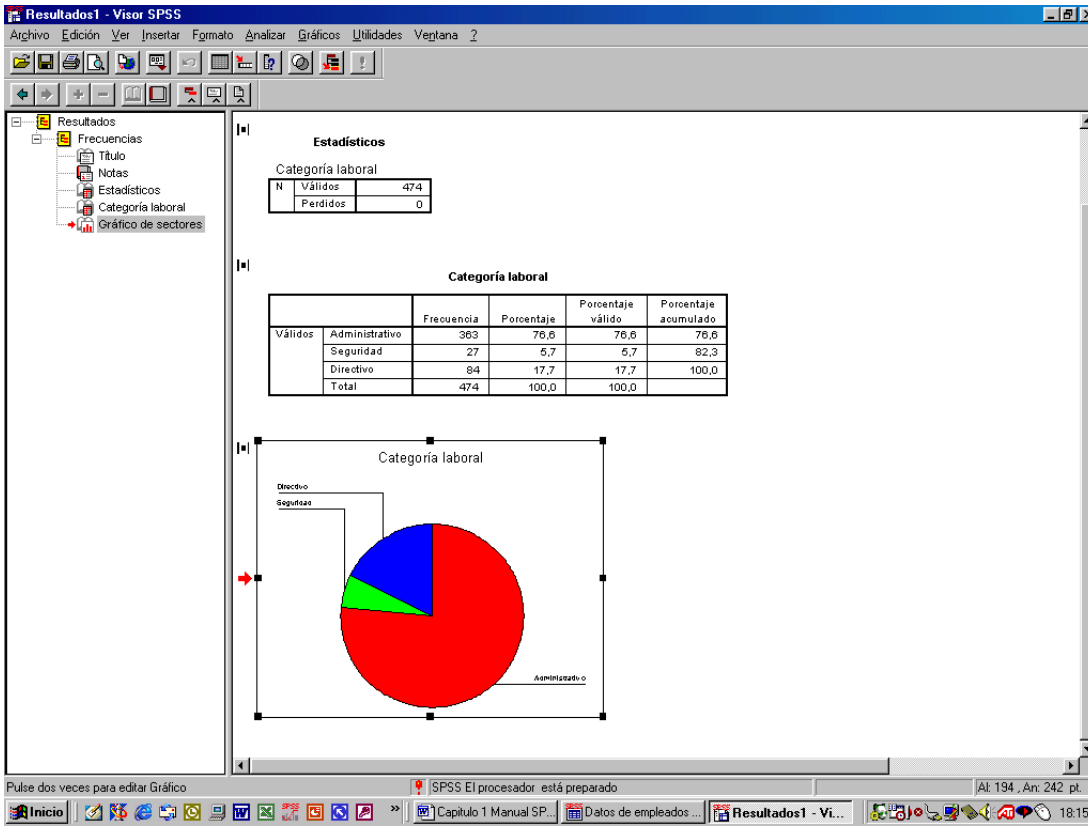


Fig. 1.10 ***Visor de SPSS*** **con algunos resultados del procedimiento** ***Frecuencia***

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

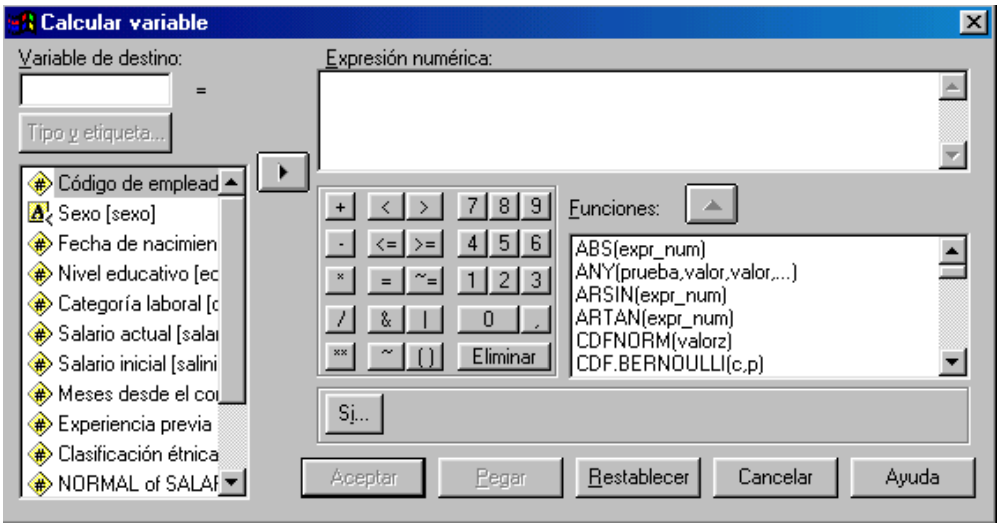


Figura 2.2 Cuadro de diálogo de creación de nuevas variables

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

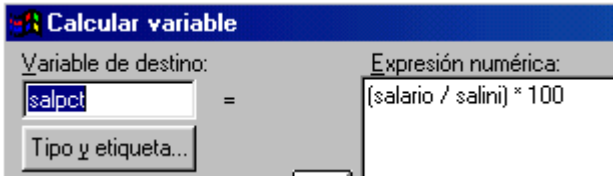


Figura 2.3 Expresión para obtener el porcentaje del salario actual respecto del inicial

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.

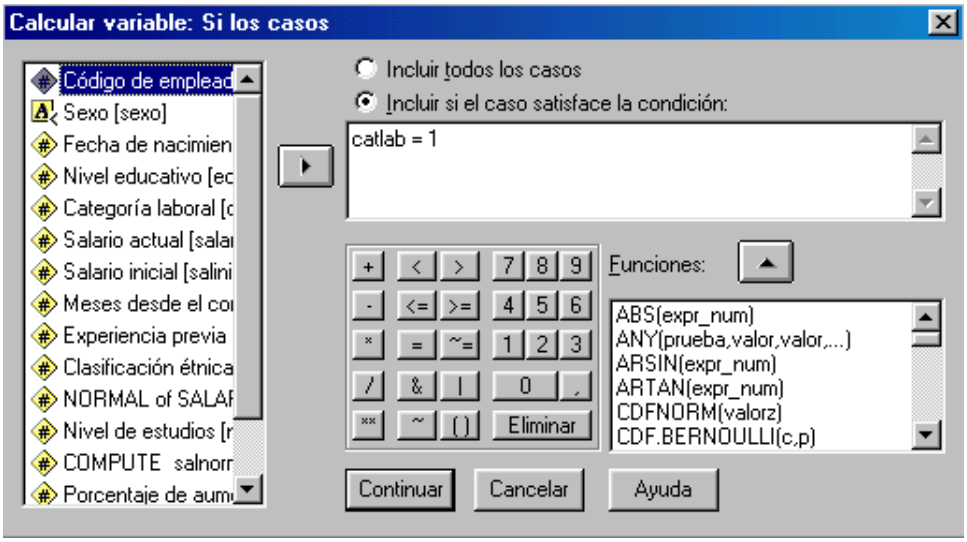


Figura 2.4 Cuadro para establecer la condición de creación de variables

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.

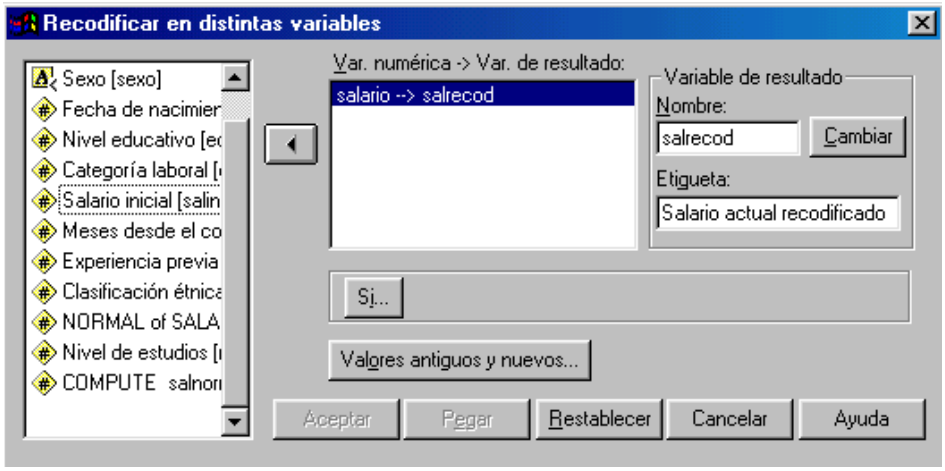


Figura 2.5 Cuadro de selección de variables a recodificar

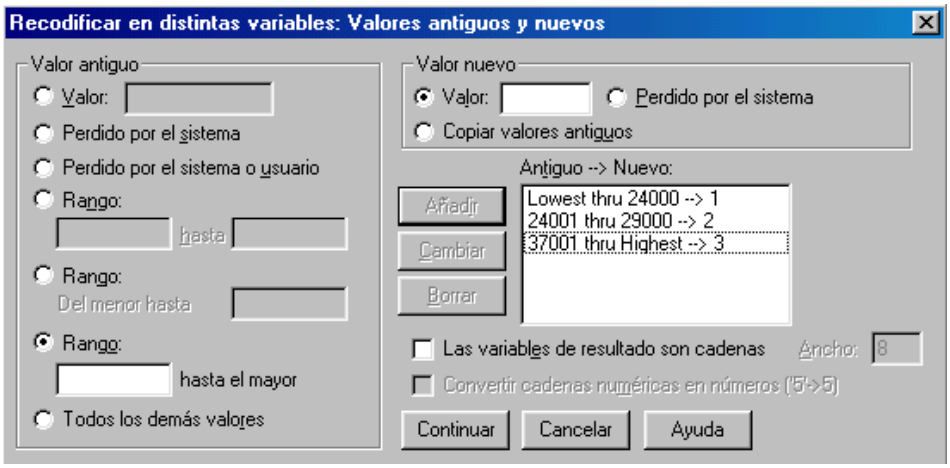


Figura 2.6 Cuadro para establecer los valores de recodificación

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...

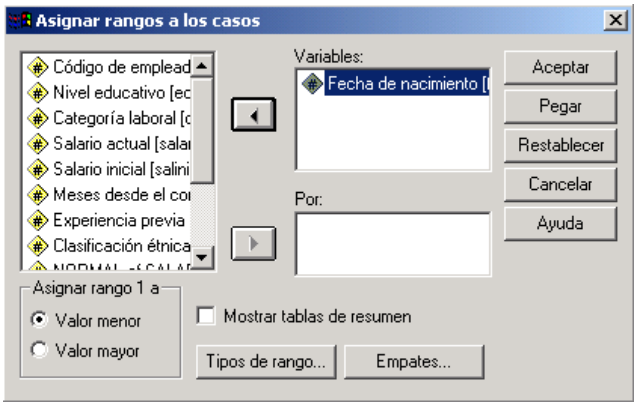


Figura 2.8 Cuadro de diálogo para asignar rangos a casos

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.

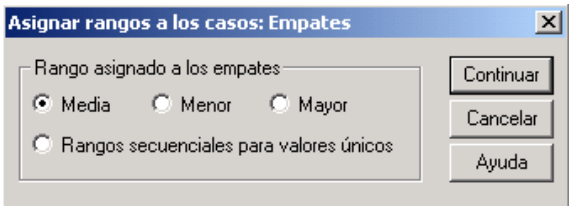


Figura 2.9(a)

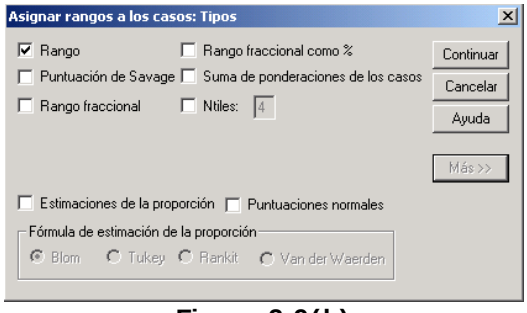
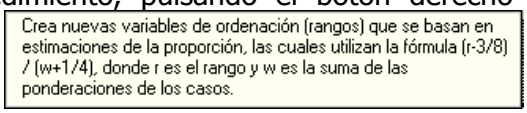


Figura 2.9(b)

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

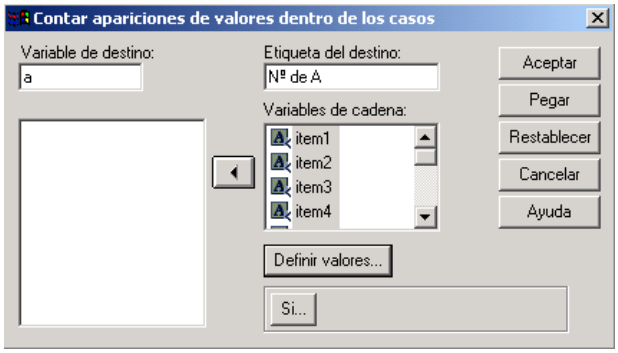


Figura 2.10(a)

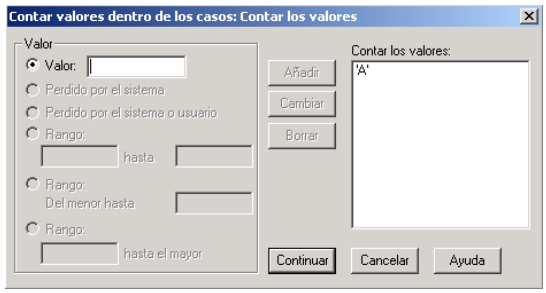


Figura 2.10(b)

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.

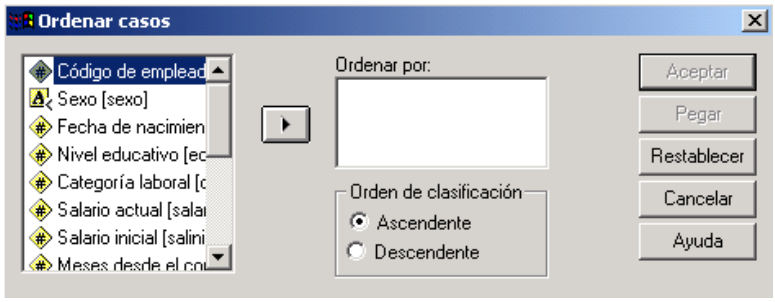


Figura 3.1 Cuadro de diálogo para ordenar casos

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

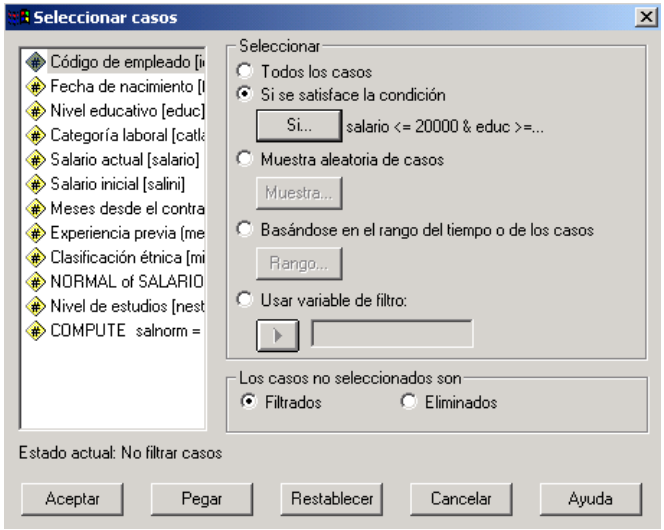


Figura 3.2 Cuadro de opciones para seleccionar casos

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

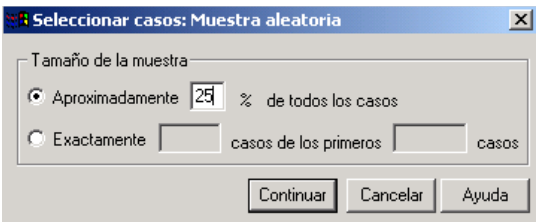


Figura 3.3 Cuadro de selección aleatoria de casos

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

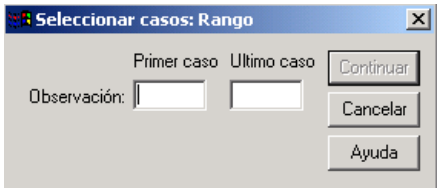


Figura 3.4 Cuadro para seleccionar un rango de casos

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...

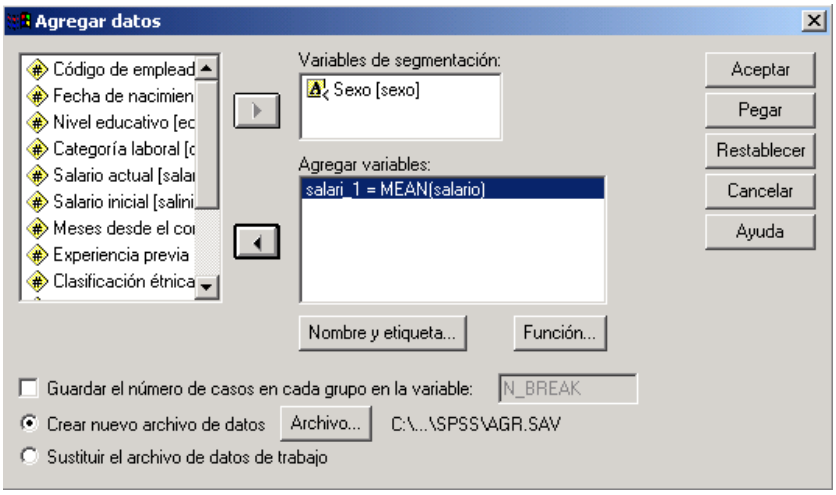


Figura 3.5 Cuadro de diálogo del procedimiento para agregar datos

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).

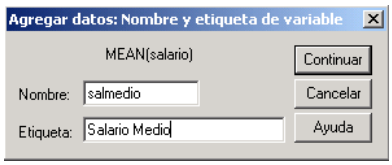


Figura 3.6 (a)

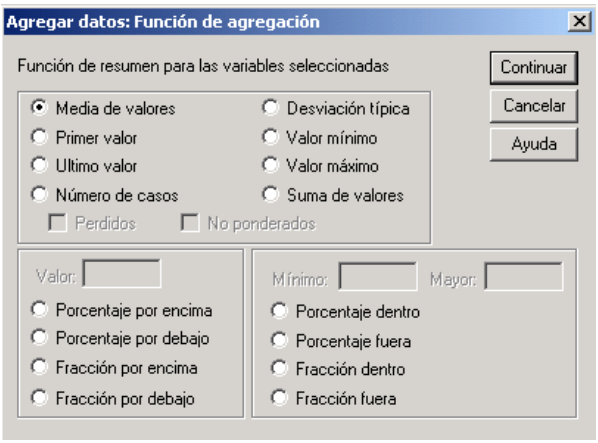


Figura 3.6 (b)

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 labora**l 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.

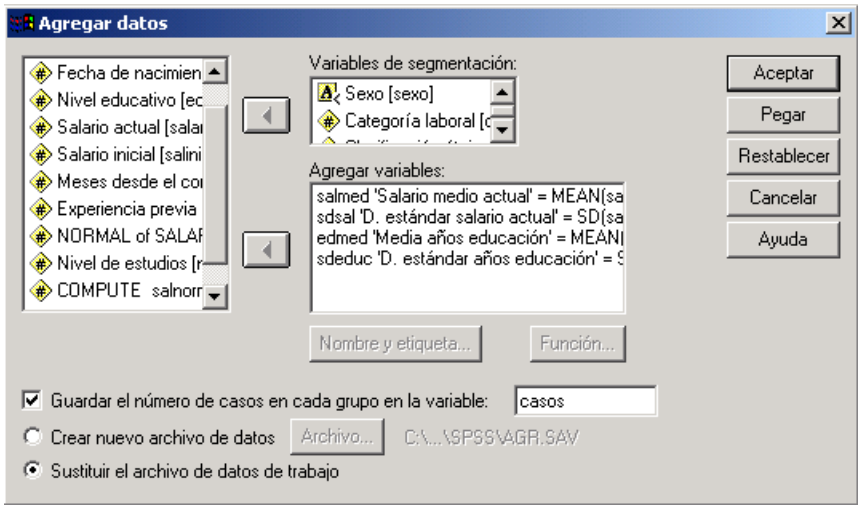


Figura 3.7 Cuadro de agregación de datos con varias variables de segmentación y varias variables agregadas.

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.

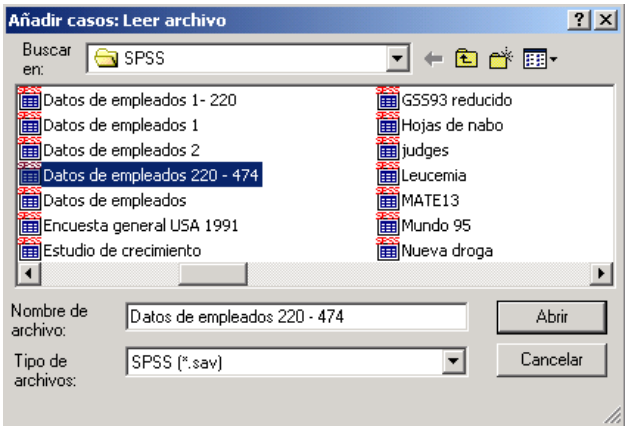


Figura 3.8. Cuadro de diálogo de Añadir casos: Leer archivo

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:

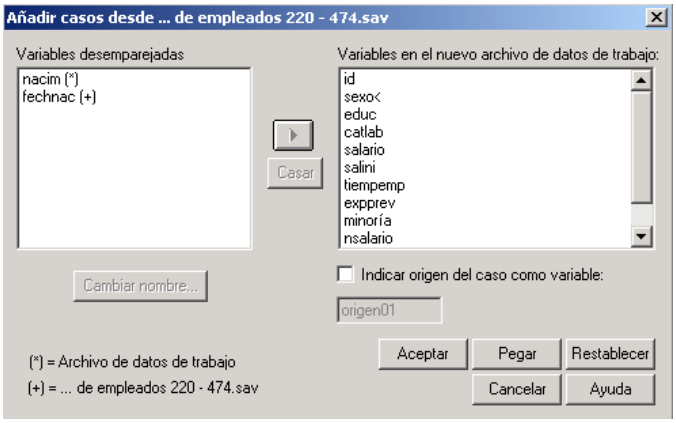


Figura 3.9 Cuadro de diálogo Añadir casos desde...

- 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.

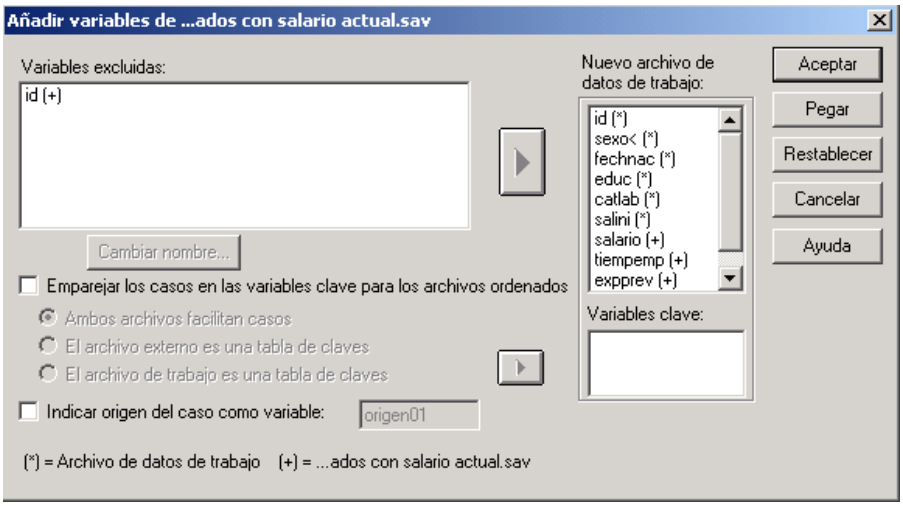


Figura 3.10. Cuadro de diálogo *Añadir variables de...*

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**.

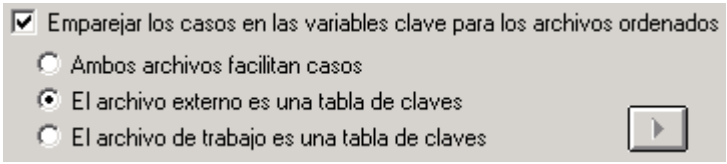
Cuando el archivo externo en vez de casos contiene una tabla de claves, el proceso es el mismo, y la única diferencia en el proceso es señalar dicha opción en el cuadro de diálogo **Añadir variables de...** El resultado del proceso de fusión es tal que cada caso del archivo externo puede ser emparejado 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**.

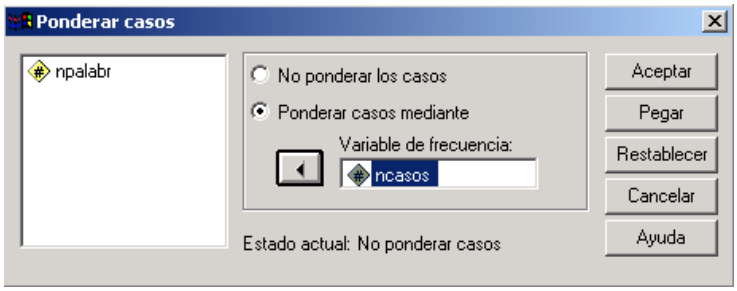


Figura 3.11. Cuadro de diálogo para *Ponderar casos*

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.

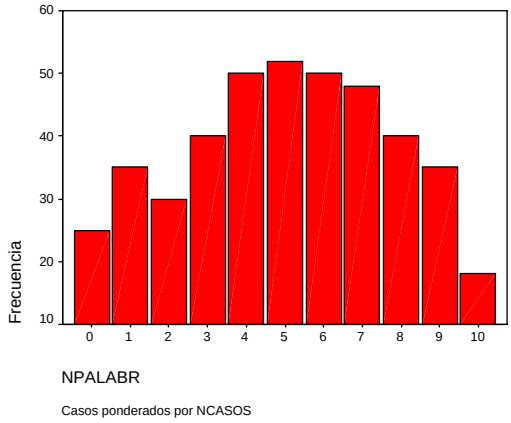


Figura 3.12 (a)

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:

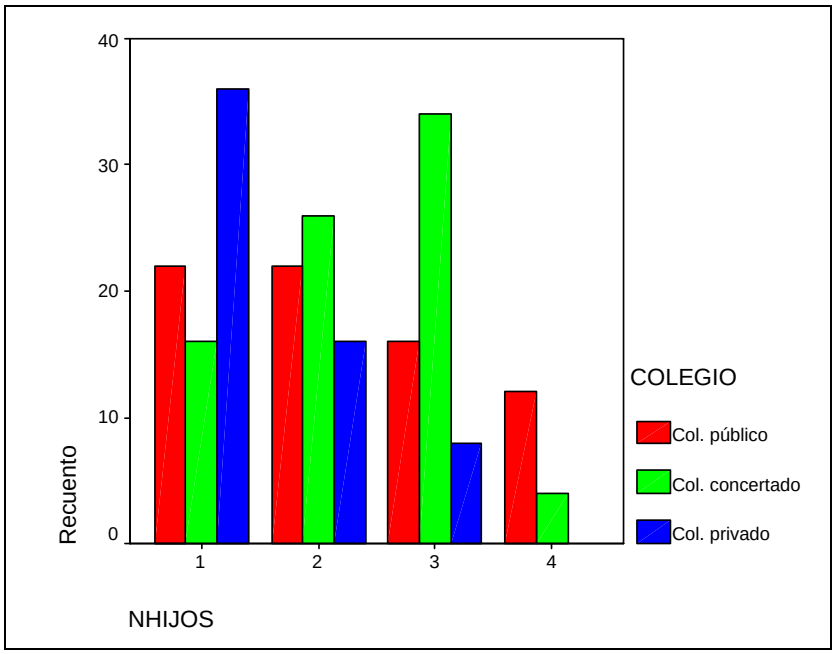
4 Aunque en la variable colegio aparecen las etiquetas (Col. público, privado, etc.) en el editor de datos se introducen los valores numéricos correspondientes a cada categoría.

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.

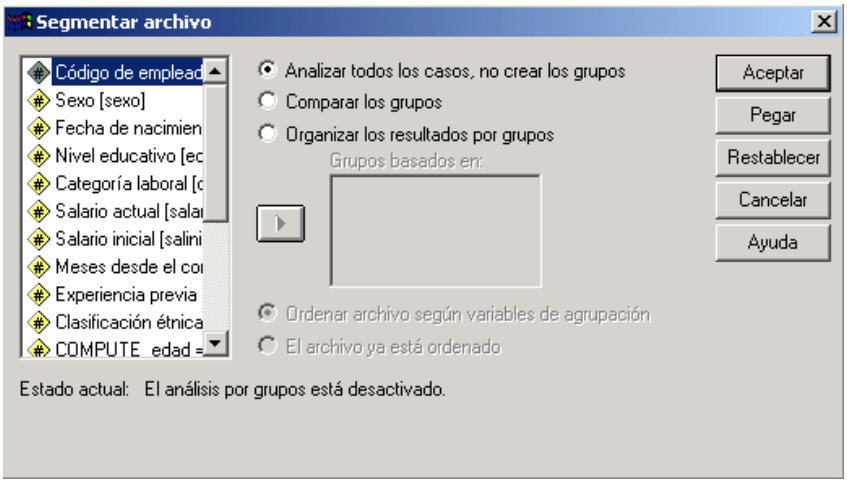


Figura 3.13 Cuadro de diálogo de Segmentar archivo

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.

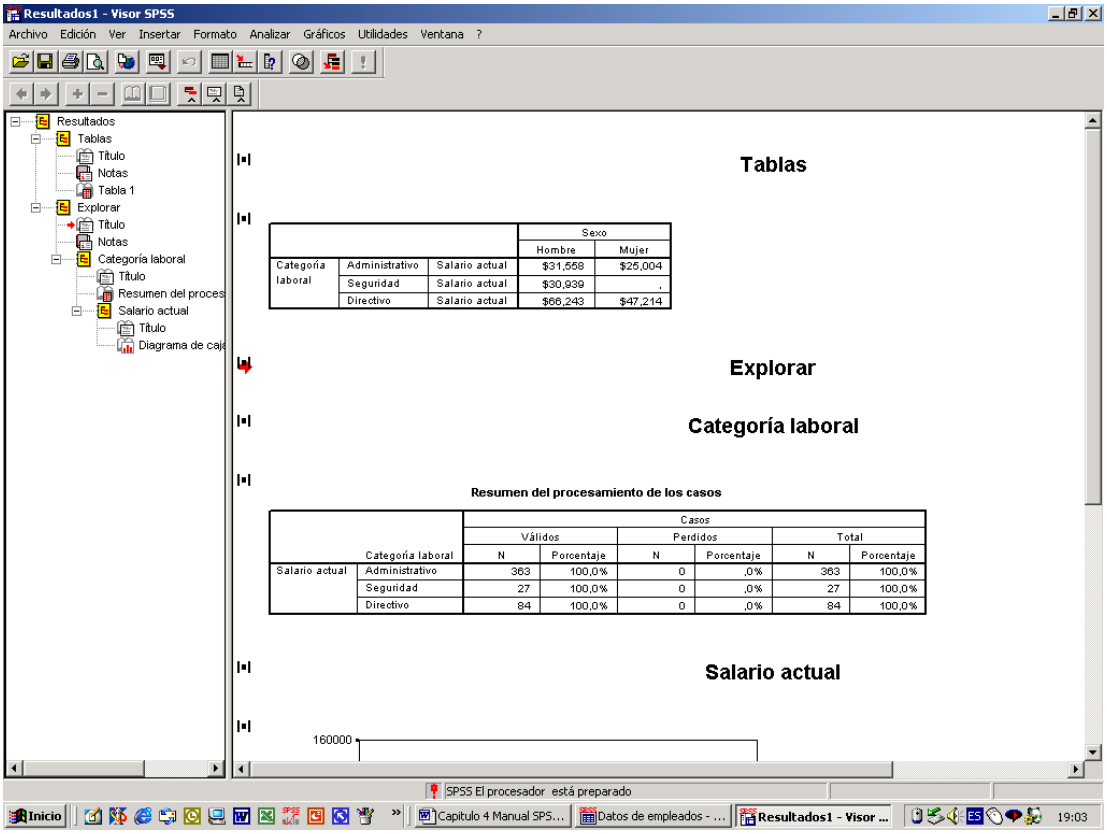


Figura 4.1 Aspecto del Visor de resultados

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.

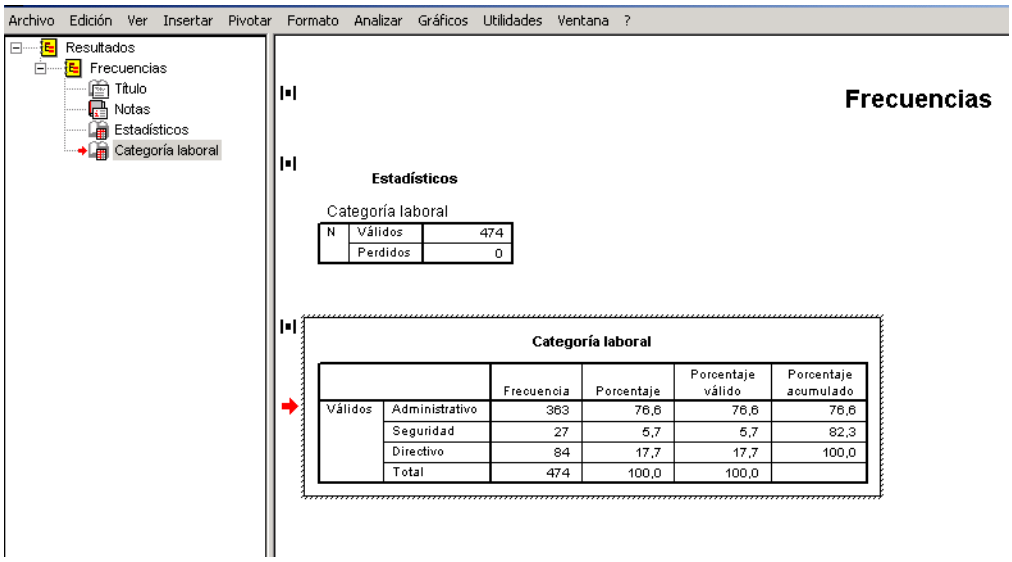


Fig. 4.2 Tabla pivote en su ventana de edición.

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.

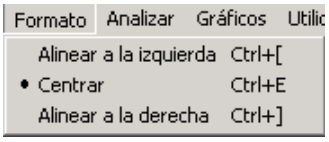


Figura 4.3(a)

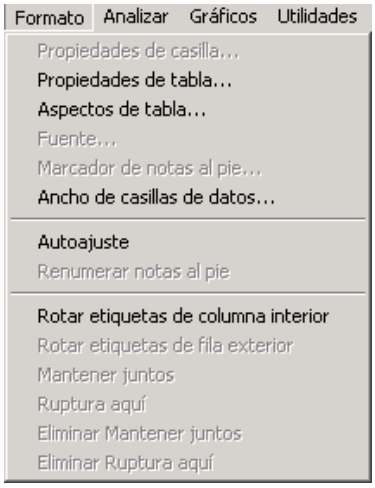


Figura 4.3 (b)

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.

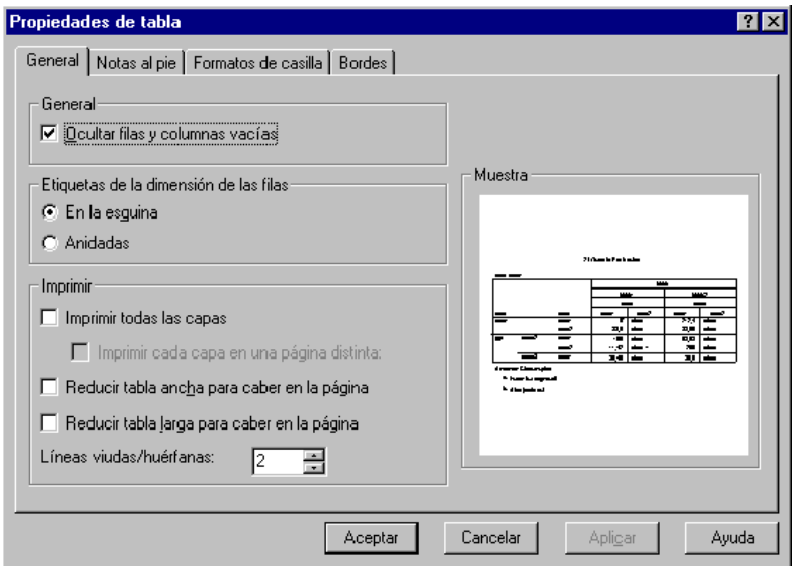


Figura 4.5 Carpeta para elegir diferentes propiedades de las tablas pivote.

- ***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**

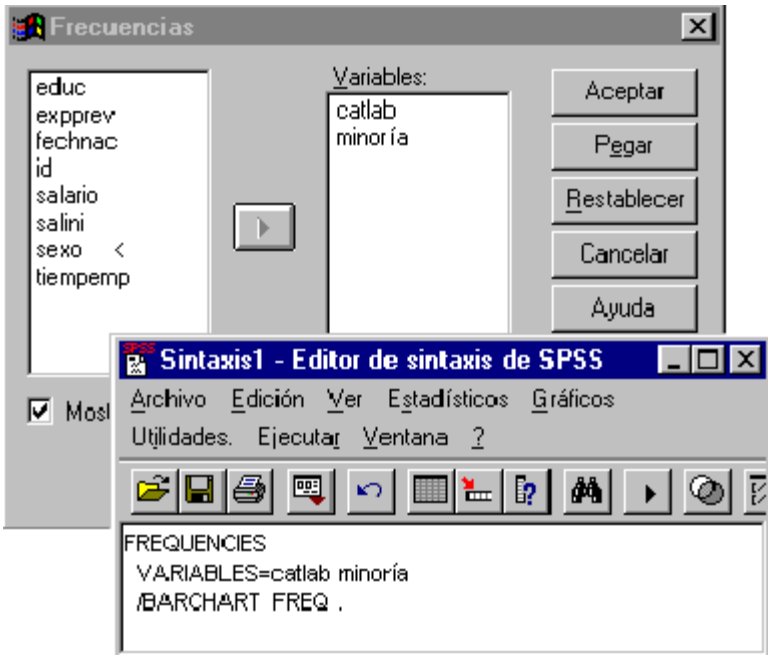


Figura 5.1 Sentencias del procedimiento Frecuencias pegado en la ventana de sintaxis

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.

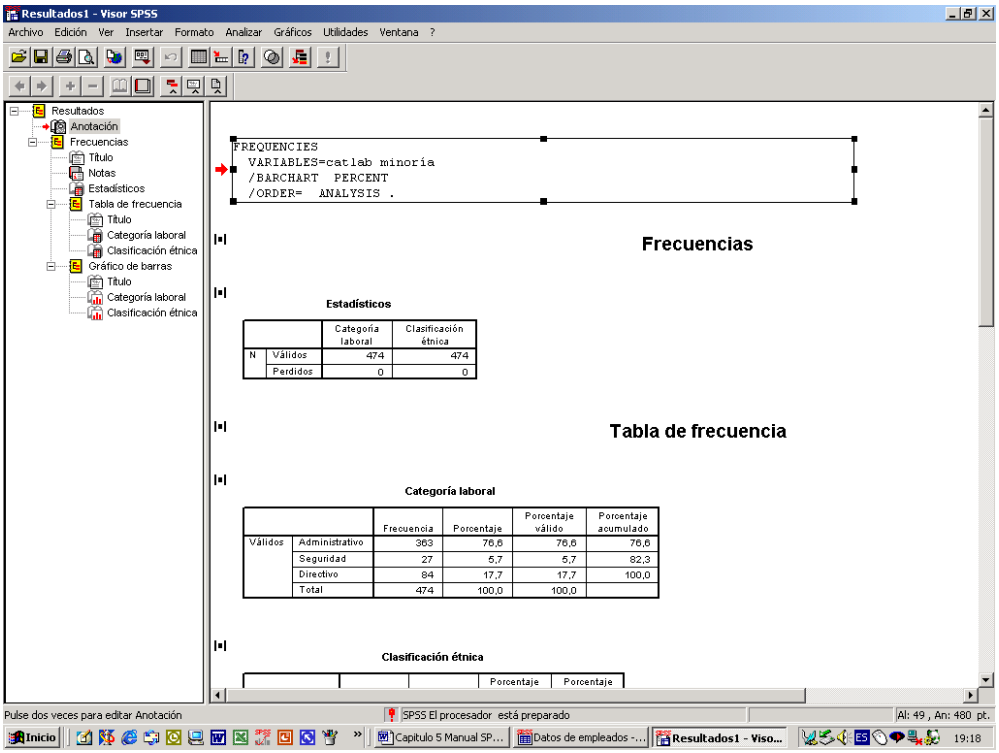


Figura 5.2 Sentencia del procedimiento Frecuencias insertadas en el Visor.

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).

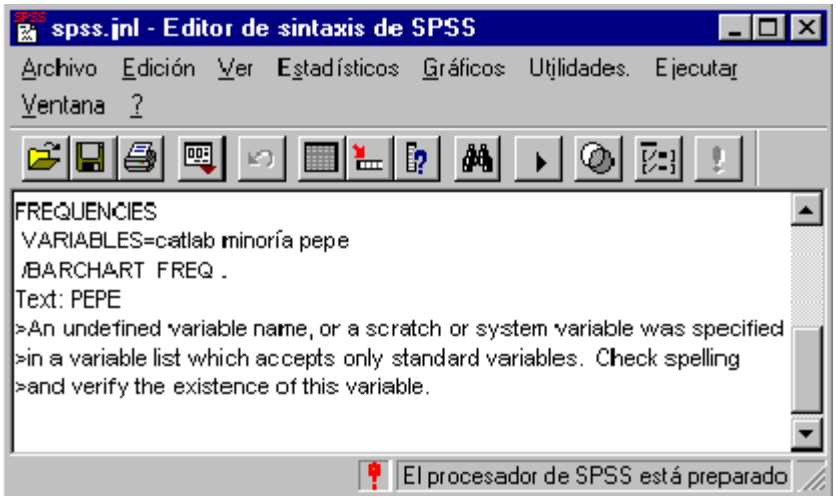


Figura 5.3 Mensaje de error insertado en el archivo de trabajo diario de SPSS

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.

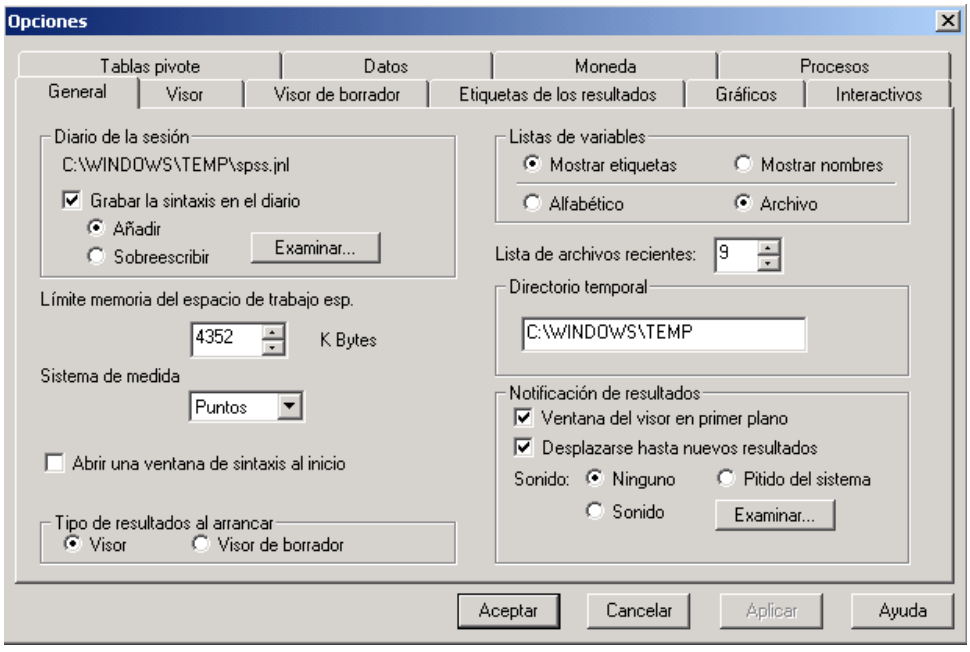


Figura 6.1 Carpetas con las diferentes opciones de SPSS

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.

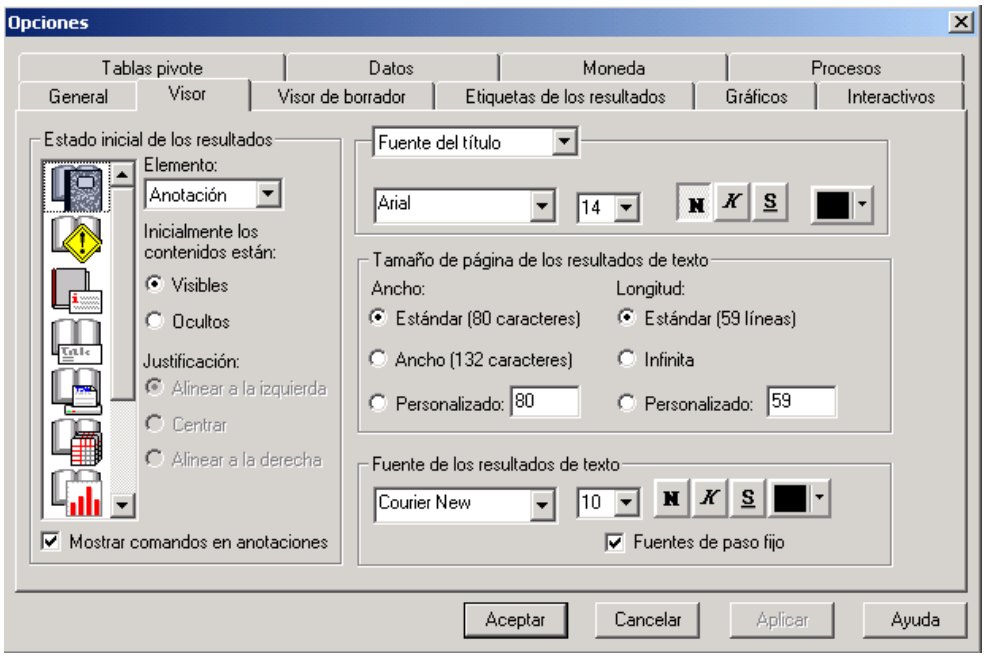


Figura 6.2 Opciones del Visor de resultados

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.

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

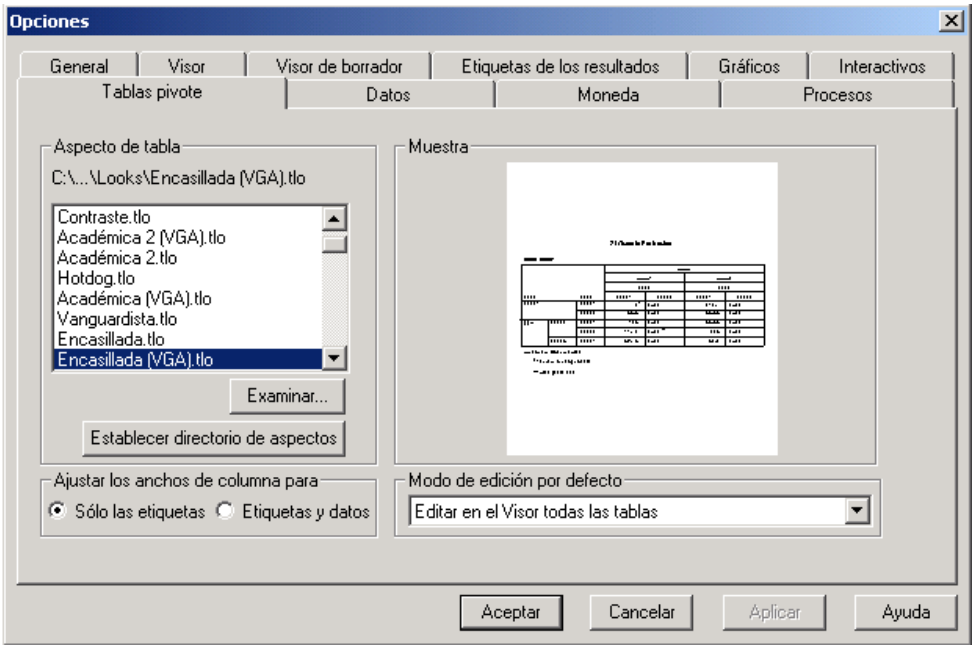


Figura 6.3 Opciones sobre el aspecto de las tablas pivote

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.

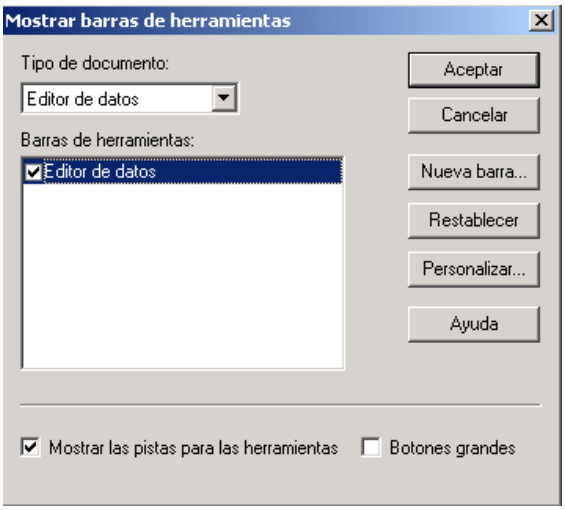


Figura 6.4 Cuadro de diálogo para acceder a las barras de herramientas de las diferentes ventanas de SPSS.

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.

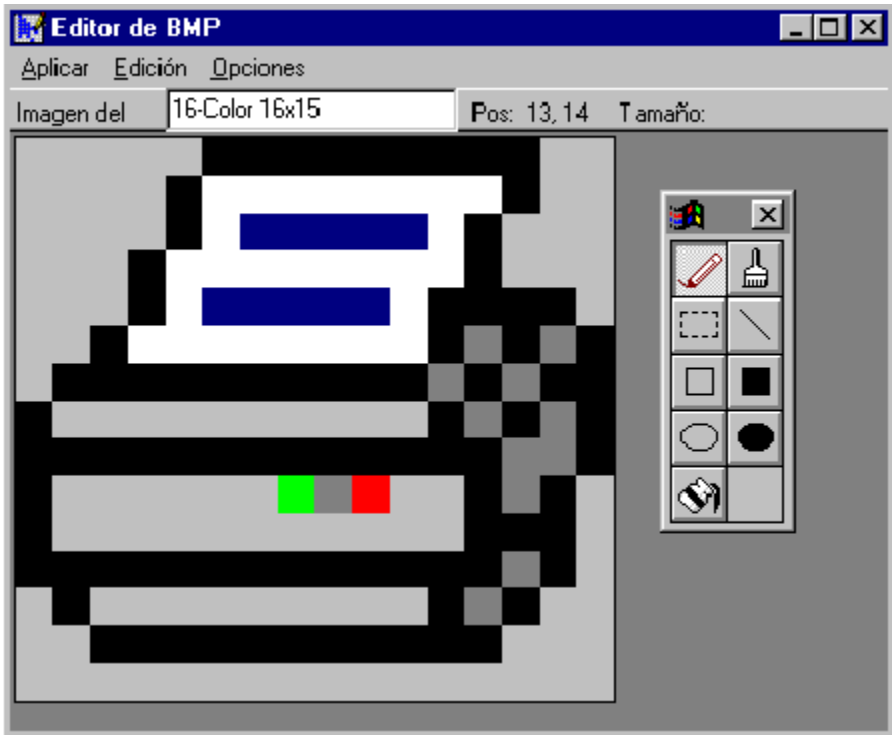


Figura 6.6 Icono de impresora en el editor de mapas de bits.

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.

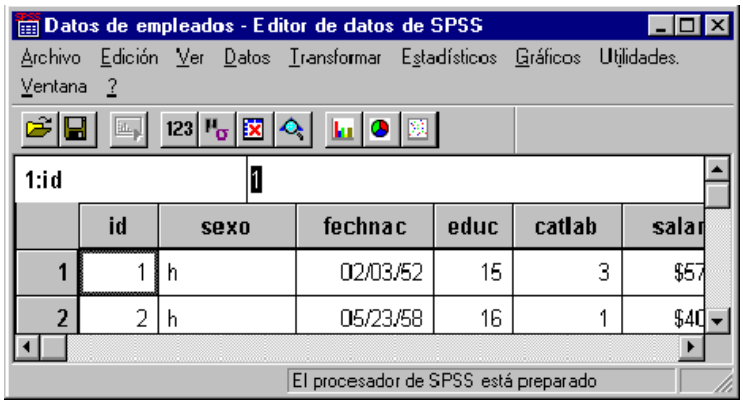


Figura 6.7 Editor de datos con iconos de funciones incorporados por el usuario

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.

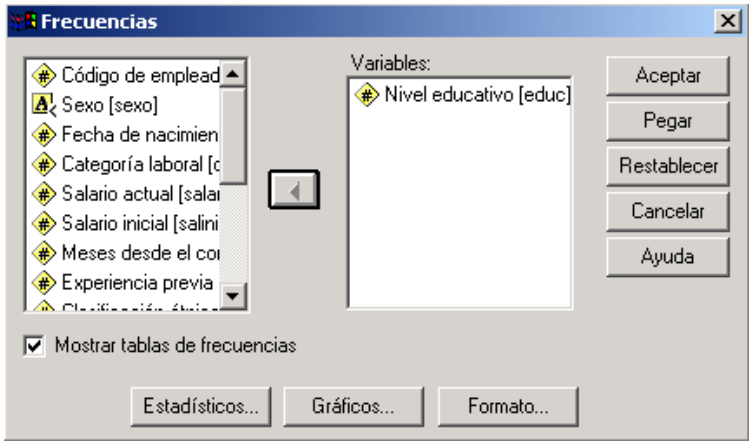


Figura 7.1 Cuadro de diálogo de *Frecuencias*

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)

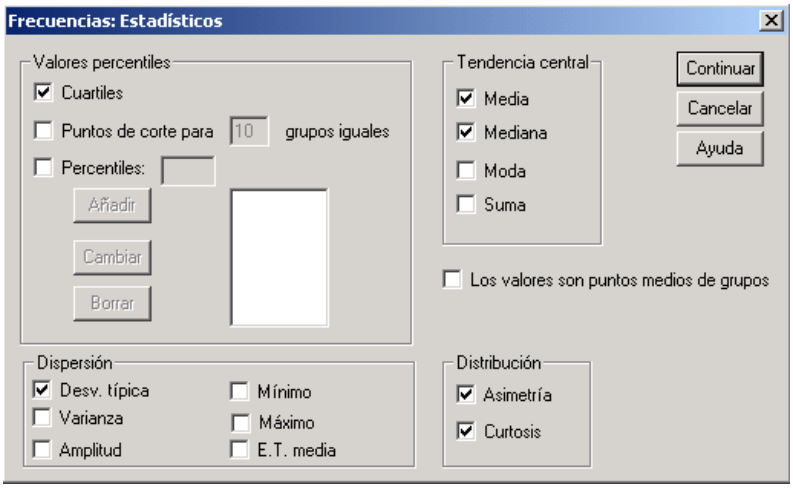


Figura 7.2 (a)

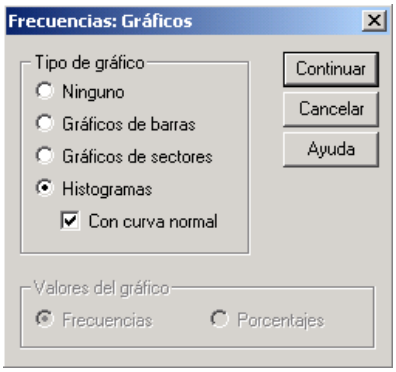


Figura 7.2 (b)

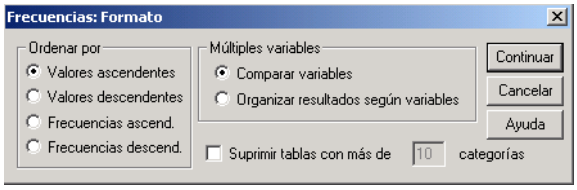


Figura 7.2 (c)

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.

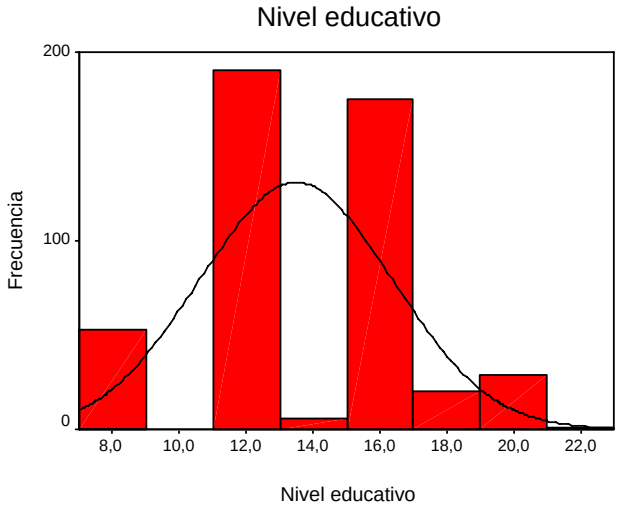


Figura 7.3 Histograma de la variable educ (nivel educativo)

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.

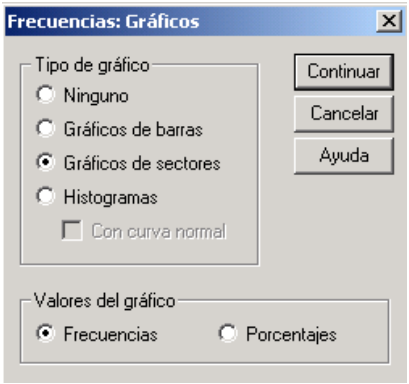


Figura 7.4 Cuadro de Gráficos de Frecuencias

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.

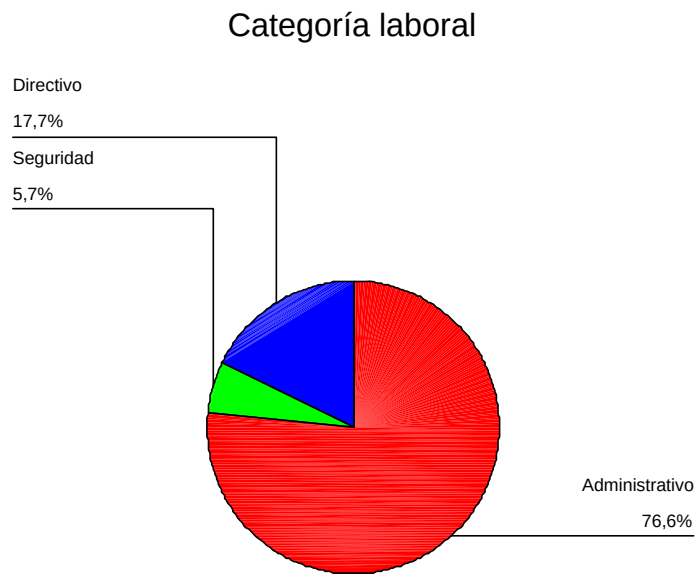


Figura 7.5 Gráfico de sectores de Categoría laboral expresado en porcentajes

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.

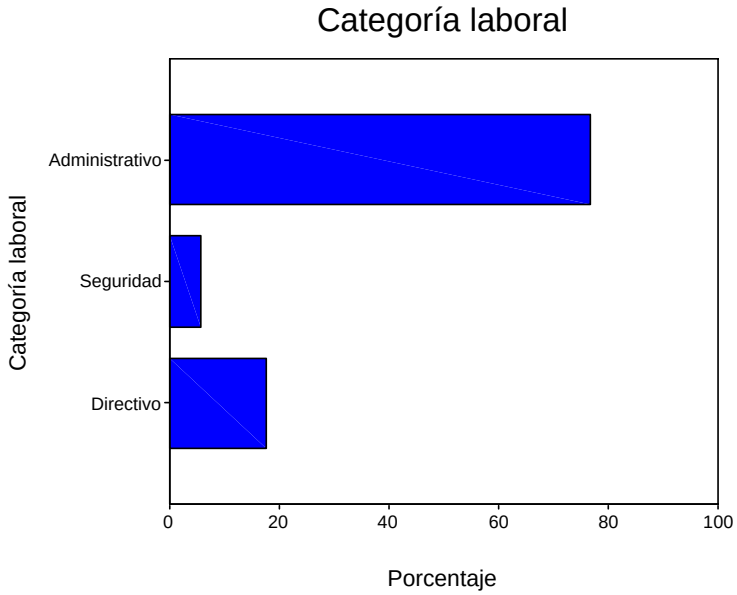


Figura 7.6 Gráfico de barras de Categoría laboral expresado en porcentajes

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.

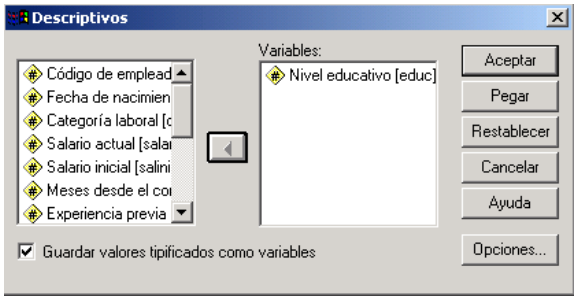


Figura 7.7 (a)

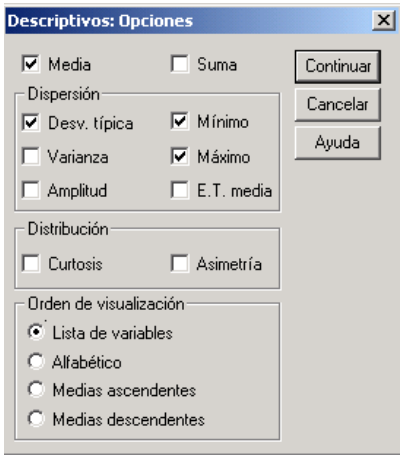


Figura 7.7 (b)

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:

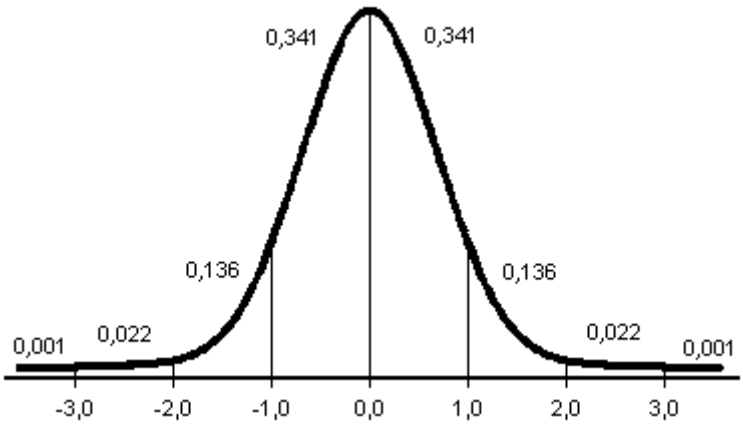


Figura 7.8 Curva normal tipificada y proporción de casos entre puntuaciones típicas

- 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

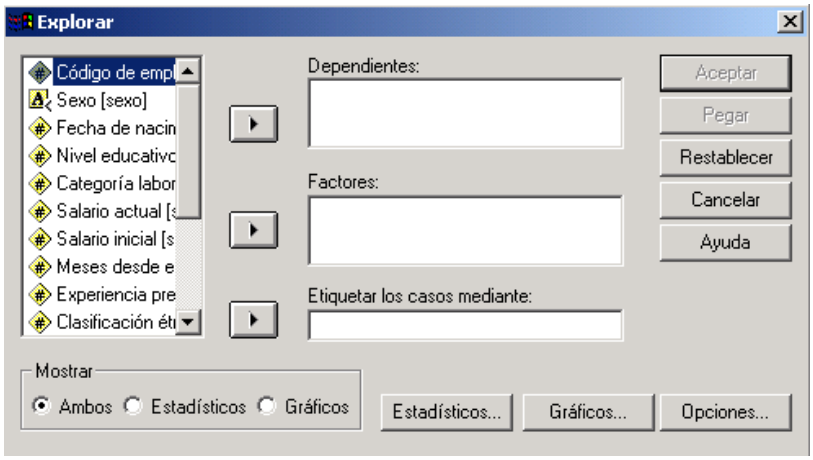


Figura 8.1 Cuadro de diálogo de Explorar

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.

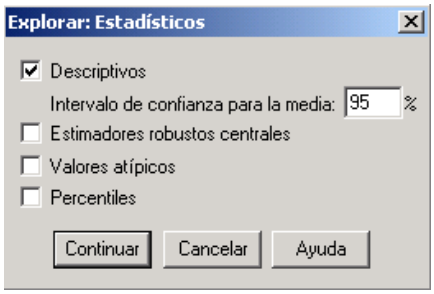


Figura 8.2 Cuadro de Estadísticos de Explorar

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.

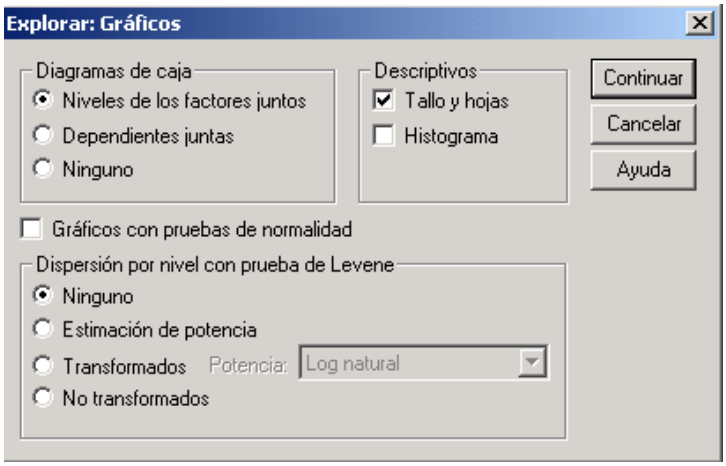


Figura 8.3. Cuadro de Gráficos de Explorar

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.

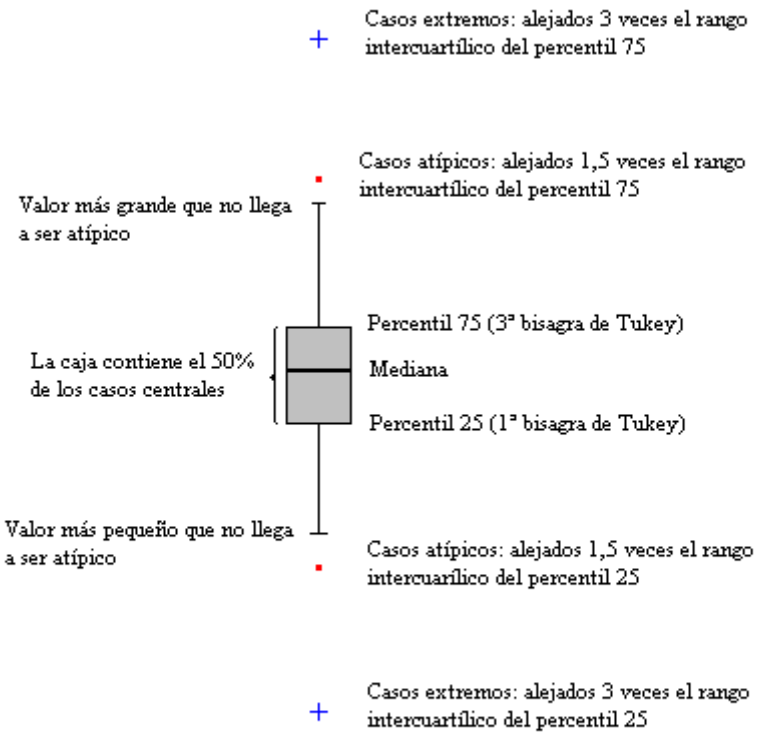


Figura 8.3 Detalles de un diagrama de caja

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 .   00000000000111111233333444
   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.

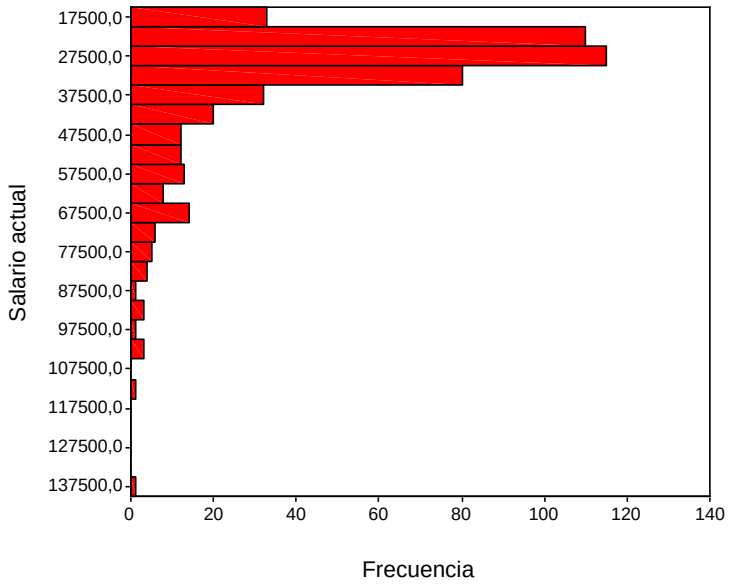


Figura 8.7 Histograma de la variable *salario actual*

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.

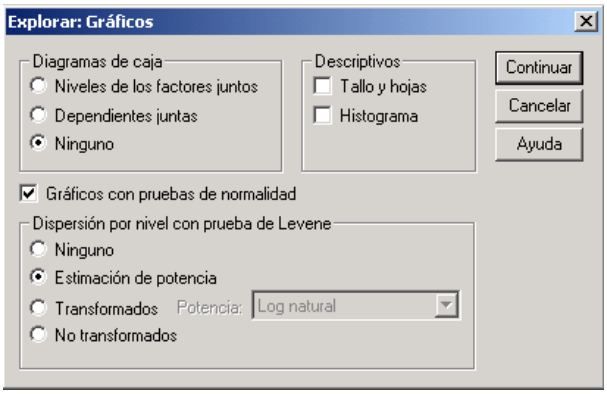


Figura 8.8 Opciones de gráficos de normalidad y de homogeneidad de varianza del cuadro Gráficos de Explorar.

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

		Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	ESTUDIO	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

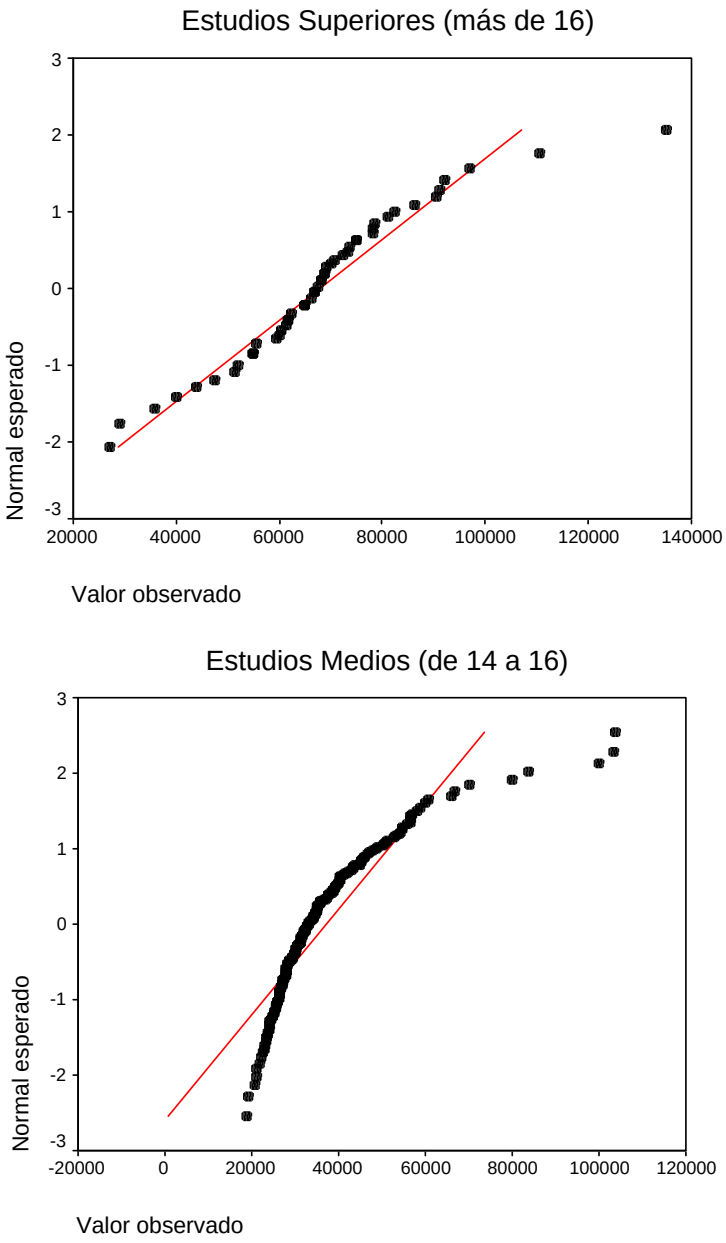


Figura 8.9 Gráficos Q-Q de normalidad con tendencias

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).

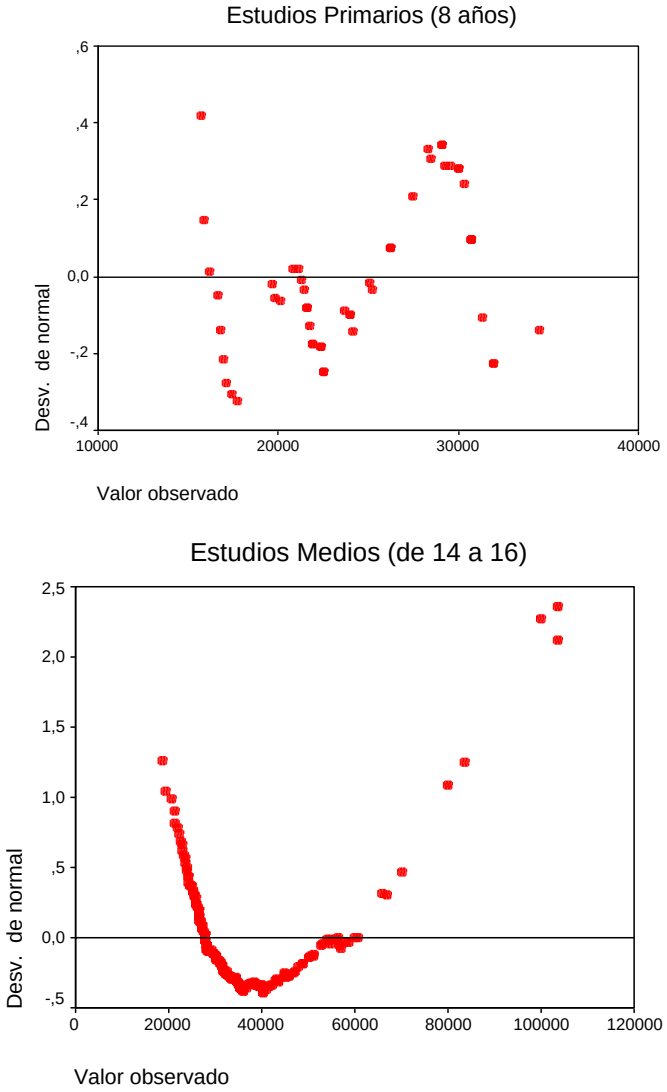


Figura 8.10 Gráficos Q-Q normal sin tendencia

8.3.2 Homogeneidad de varianzas

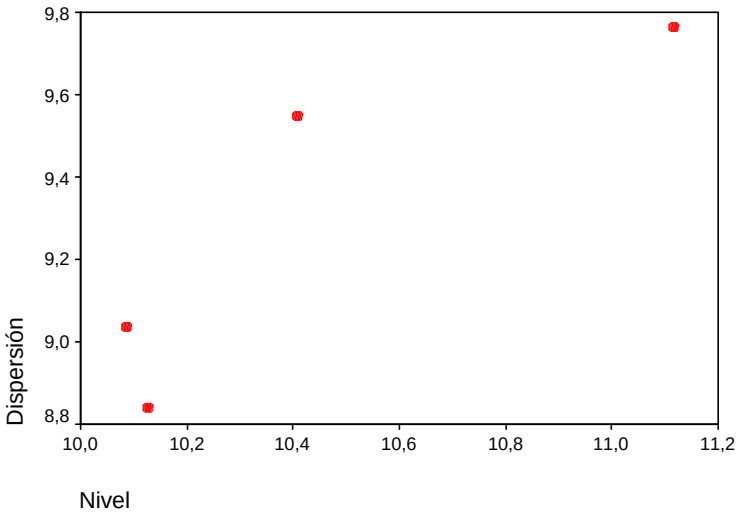
Como ya se ha señalado, el procedimiento ***Explorar*** también informa de si las varianzas son o no homogéneas. En el caso del salario actual se puede ver que no hay homocedasticidad entre los grupo de estudio, tal como informa el Estadístico de *Levene* que se puede ver en la Tabla 8.3

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.

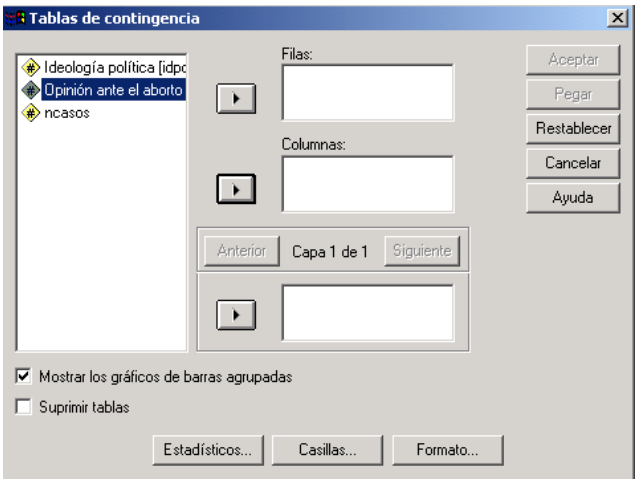


Figura 9.1 Cuadro de diálogo de Tablas de contingencia

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.

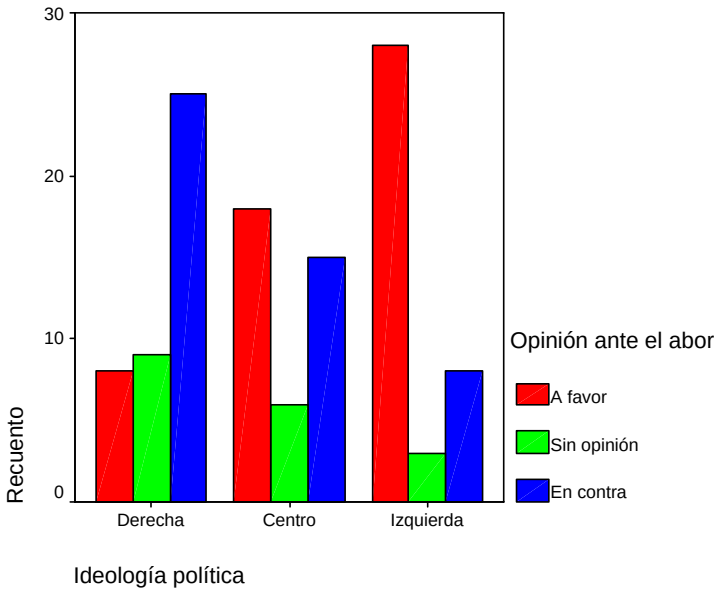


Figura 9.2 Gráfico de barras agrupadas de *ideología pol´tica* y *opinión ante el aborto*

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)

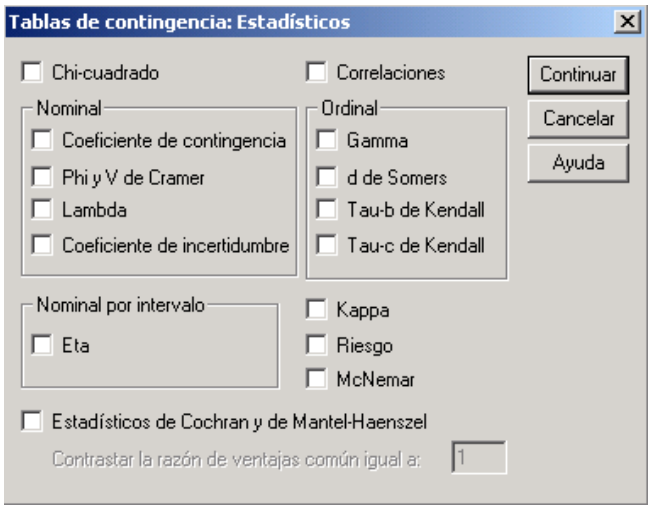


Figura 9.3 Cuadro de estadísticos de Tablas de contingencia

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		
		Sin problemas	Con problemas	Total
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 π *= 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.

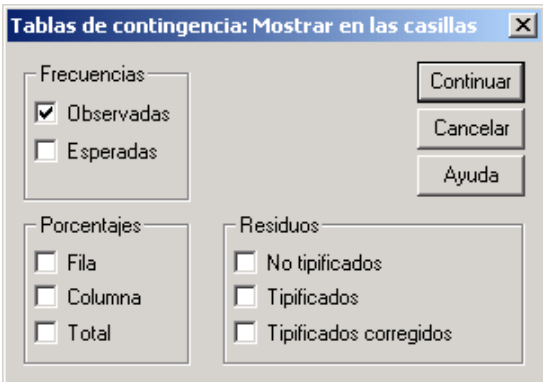


Figura 9.4 Contenido de las casillas en las Tablas de contingencia

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.

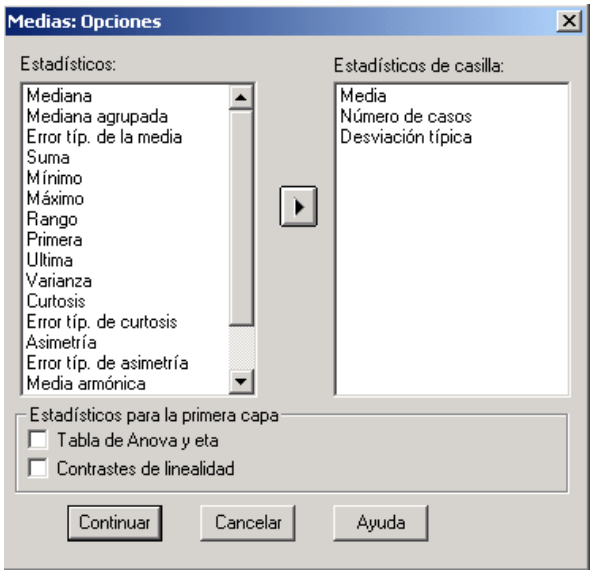


Figura 10.2 Cuadro de diálogo *Opciones* de *Medias*

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.

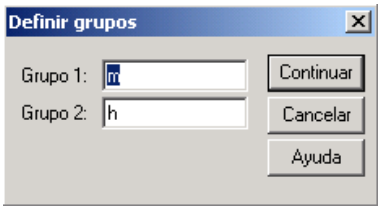


Figura 10.5 Cuadro para Definir grupos

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

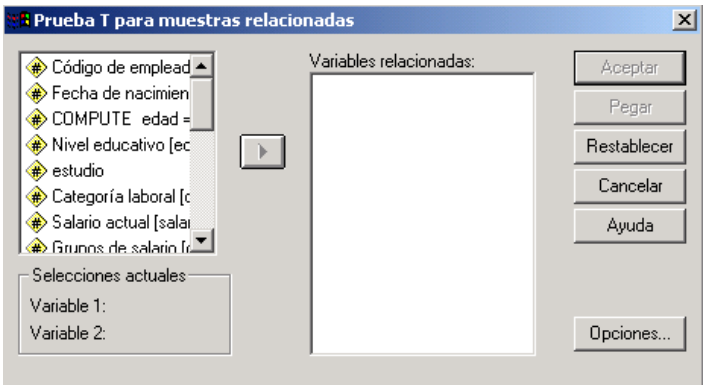


Figura 10.6 Cuadro de diálogo de Prueba T para dos muestras relacionadas

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

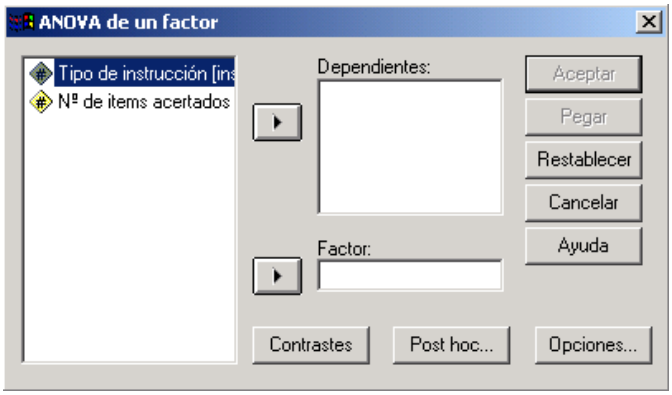


Figura 11.2 Cuadro de diálogo ANOVA de un factor

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

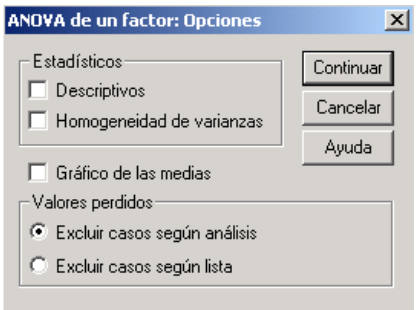


Figura 11.3 Cuadro de opciones de ANOVA de un factor

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.

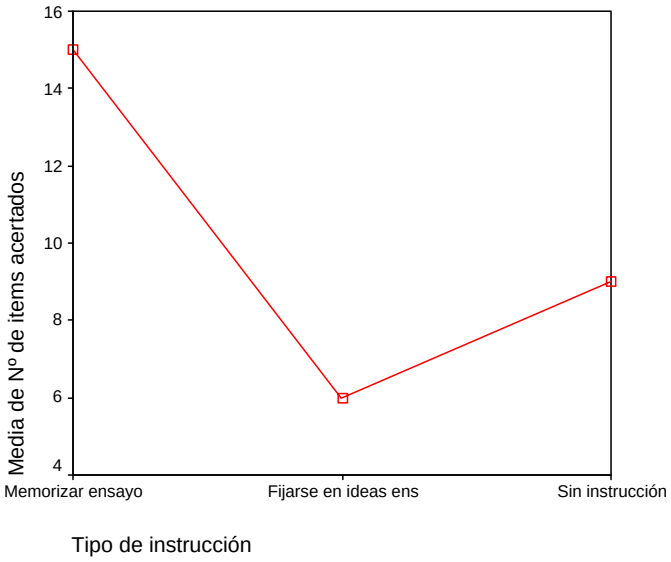


Figura 11.4 Gráfico de medias del procedimiento ANOVA

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.

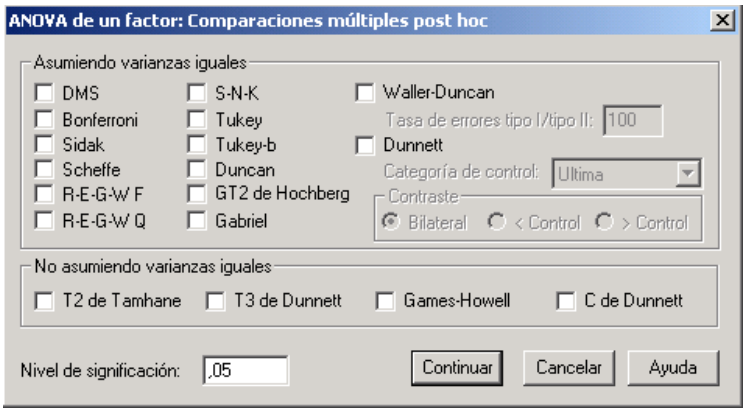


Figura 11.5 Cuadro de comparaciones múltiples de ANOVA

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.

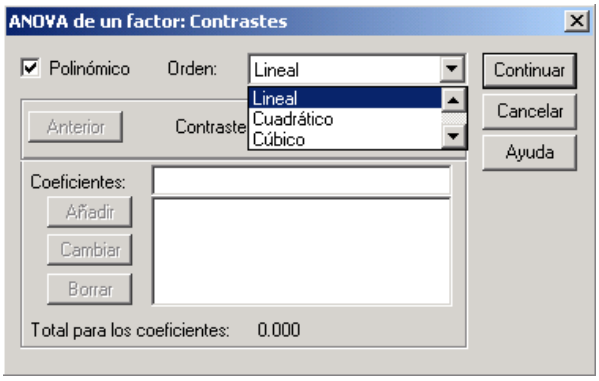


Figura 11.6 cuadro de diálogo de *Contrastes* de *ANOVA de un factor*

- **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.

12. El Modelo Lineal General.

Análisis de varianza factorial Univariante.

12.1 Introducción

En el capítulo anterior se ha visto, dentro de los varios procedimientos que permiten la comparación de medias, el ANOVA de un factor, que permitía contrastar la hipótesis de igualdad de medias de las poblaciones definidas por los diferentes niveles en que se podía segmentar el factor o variable independiente (VI). Cuando se desea estudiar el efecto de más de un factor sobre la variable dependiente (VD) es preciso recurrir a los modelos factoriales de análisis de varianza que permiten estudiar el efecto de diversos factores, tanto de manera individual como conjunta.

Cuando sólo se tiene en cuenta un factor, se estudia su efecto sobre la VD y se especifican diversos contrastes entre los niveles del factor, si el resultado del ANOVA es significativo. Sin embargo, cuando en el estudio intervienen dos factores, hay tres efectos que deben considerarse: los efectos de cada factor por separado sobre la VD, que se conocen como *efectos principales*, y el *efecto de la interacción* de ambos factores sobre la VD. Si el número de factores fuera tres, los efectos a estudiar serían 7 (tres principales, 3 interacciones dobles y 1 interacción triple). Si el número de factores fueran cuatro, los efectos a estudiar serían 15 (4 principales, 6 interacciones binarias, 4 interacciones triples, y 1 interacción cuádruple), y así sucesivamente.

Además del estudio de los efectos sobre la VD de varios factores, el procedimiento ***Univariante*** permite realizar análisis de covarianza y análisis de regresión, así como estudiar modelos aleatorizados en bloques y modelos jerárquicos con factores anidados. En este capítulo, sólo vamos a considerar el modelo factorial de dos factores, conocido en la jerga del diseño como diseño factorial completamente aleatorizado, y el análisis de covarianza de un factor, conocido como diseño factorial con un factor y una variable de control o concomitante.

12.2 El diseño factorial completamente aleatorizado

Como ya se ha señalado, en estos diseños se exploran los efectos que cada factor tiene sobre la VD y los efectos de la interacción. La hipótesis nula para cada factor dice que las medias de las poblaciones definidas por los grupos o niveles del factor son iguales. Por su parte, las hipótesis referidas a las interacciones afirman que éstas no existen. Para el contraste de estas hipótesis se utiliza el estadístico F, y según sea su valor crítico se aceptará o no la hipótesis planteada.

Los supuestos para que el análisis de varianza factorial pueda realizarse, son los ya expuestos en el capítulo anterior. Cuando hay un solo factor, el número de poblaciones involucradas son tantas como niveles del factor, y cuando hay más de

un factor, el número de poblaciones involucradas serán tantas como el producto de los niveles de cada uno de los factores. Si, por ejemplo, ser realiza el análisis de varianza factorial con dos factores, el primer con 3 niveles, y el segundo con 4 niveles, el número total de poblaciones será 12 (3 x 4 = 12). Estas 12 poblaciones deben ser normales y homocedásticas.

Para explicar el proceso de análisis vamos a utilizar los datos de una investigación hipotética.

> *"Se desea estudiar el efecto del nivel de privación alimenticia (Factor A) y ciertas drogas (Factor B) sobre la ejecución de determinadas tareas en monos. Para ello se presenta a los monos 3 objetos (2 iguales y uno diferente) y su tarea consiste en aprender a seleccionar el objeto no duplicado. Al lado del objeto correcto se sitúa un recipiente con comida. En cada ensayo se presentan los tres objetos y el mono debe elegir uno de ellos. La VD es el número de errores en 20 ensayos. El factor A tiene dos niveles de privación: 1 hora de privación (nivel 1); y 24 horas de privación (nivel 2). El factor B tiene tres niveles: un grupo de control sin droga (nivel 1); un grupo con un tipo de droga X (nivel 2) ; un grupo con un tipo de droga Y (nivel 3)"*

Los datos del experimento se recogen en la siguiente tabla:

A1	A1	A1		A2	A2	A2
B1	B2	B3		B1	B2	B3
1	13	9		15	6	14
4	5	16		6	18	7
0	7	18		10	9	6
7	15	13		13	15	13

Para realizar el análisis factorial con más de un factor y una sola VD se sigue la secuencia

Analizar → Modelo Lineal General → Univariante...

y se accede al cuadro de diálogo de la Figura 12.1.

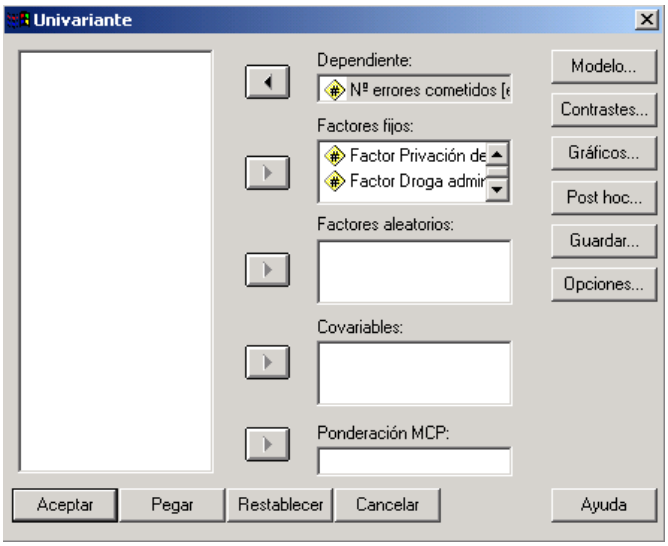


Figura 12.1 Cuadro de diálogo de *Univariante*

Las variables dependientes se seleccionan del listado de variables de la izquierda. La variable debe estar medida a nivel de escala, es decir de tipo numérico.

Los factores (variables categóricas) se seleccionan del listado de variables de la izquierda y se incorporan a la ventana correspondiente según el carácter del factor. Un **factor de efectos fijos** es aquel en que el investigador establece los niveles del factor (por ejemplo, administrar una droga u otra o no administrarla) o viene determinada por la propia naturaleza de la variable (sexo, estado civil, etc.). *Los niveles concretos que toma un factor de efectos fijos constituyen la población de niveles sobre los que se realiza la inferencia*.

Un **factor aleatorio** es aquel cuyos niveles se seleccionan de forma aleatoria entre todos los posibles niveles del factor (por ejemplo, administrar determinadas cantidades de droga seleccionadas de entre un intervalo de cantidades de droga). *Los niveles concretos que toma un factor de efectos aleatorios es sólo una muestra de la población de niveles sobre los que se hace la inferencia*.

Como en el ejemplo los factores son fijos, se pasan las variables a las listas correspondientes, y el resultado del análisis con las opciones por defecto se muestran en la Tabla 12.1.

Tabla 12.1 Factores inter-sujetos y Prueba de sus efectos

Factores inter-sujetos

		Etiqueta del valor	N
Factor Privación de comida	1	1 hora	12
	2	24 horas	12
Factor Droga administrada	1	Control	8
	2	Droga X	8
	3	Droga Y	8

Pruebas de los efectos inter-sujetos

Variable dependiente: Nº errores cometidos

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	280,000[a]	5	56,000	3,055	,036
Intersección	2400,000	1	2400,000	130,909	,000
PRIVAR	24,000	1	24,000	1,309	,268
DROGAS	112,000	2	56,000	3,055	,072
PRIVAR * DROGAS	144,000	2	72,000	3,927	,038
Error	330,000	18	18,333		
Total	3010,000	24			
Total corregida	610,000	23			

a. R cuadrado = ,459 (R cuadrado corregida = ,309)

La tabla sigue la estructura del modelo lineal general para un diseño factorial de dos factores con interacción, según el cual cada puntuación Y_{ijk} puede expresarse así:

$$Y_{ijk} = \mu_T + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

donde:

- μ_T = media global de la población
- α_i = promedio del efecto del tratamiento en el nivel a_i ($\alpha_i = \mu_i - \mu_T$)
- β_j = promedio del efecto del tratamiento en el nivel b_j ($\beta_i = \mu_j - \mu_T$)
- $(\alpha\beta)_{ij}$ = efecto de la interacción en la celda $a_i b_j$ ($(\alpha\beta)_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu_T$)
- ε_{ij} = error experimental asociado a cada puntuación ($\varepsilon_{ij} = Y_{ijk} - \mu_{ij}$)

La fuente *Modelo corregido* (280) recoge todos los efectos del modelo tomados conjuntamente (el de los factores, y el de la interacción: 280 = 24 + 112 + 144). La significación de esta fuente, inferior a 0,05, indica que el modelo explica una parte de la variabilidad de la VD, en concreto el 45,9%, que es el resultado de dividir la suma de cuadrados del *Modelo corregido* entre la suma de cuadrados *Total corregida* (280/610 = 0,459).

La fila *Intersección* se refiere a la constante del modelo, y surge como el producto del número de casos por el cuadrado de la media total de la VD. Esta constante permitiría contrastar la hipótesis, caso de ser de interés, de que la media total de la población es igual a cero, hipótesis que en nuestro caso se rechaza (p = 0,000 < 0,05).

Las dos siguientes filas recoge los *efectos principales* de cada uno de los factores incluidos en el modelo. El factor Privación de comida (p = 0,268 > 0,05) no es significativo. Tampoco lo es el factor Drogas (p = 0,072 > 0,05).

La siguiente fila recoge el efecto de la *interacción* entre los dos factores, que en nuestro caso sí resulta significativo (p = 0,038 < 0,05), lo que supone que el efecto de las drogas actúa en interacción con alguno de los niveles del otro factor. Posteriormente, veremos gráficamente este efecto.

La fila *Error* recoge el error residual y su media cuadrática (18,333) es una estimación insesgada de la varianza de las 6 (2x3) poblaciones estudiadas, que se supone igual en todas ellas.

La penúltima fila, *Total*, recoge la suma de los cuadrados (Y^2) de las puntuaciones de la VD, y sus grados de libertad coinciden con el tamaño de la muestra (24 sujetos).

Y, por último, la fila *Total corregida* recoge la variabilidad de la VD, es decir, la suma de todas las puntuaciones diferenciales elevadas al cuadrado, o lo que es lo mismo, la variación debida a cada efecto (los principales y la interacción y el efecto error).

$$610 = 24 + 112 + 144 + 330$$

Junto con estas relaciones, el lector puede ver que el valor de *Total* (3010) es igual a la suma del *Modelo corregido* (280) más la *Intersección* (2400) más el *Error* (330). Y, por último, el *Total corregido* (610) es la diferencia entre el *Total* (3010) y la *Intersección* (2400).

Si alguno de los efectos resulta ser significativo, se procede a efectuar los comparaciones dos a dos entre los diferentes niveles de los factores, utilizando para ello los procedimientos de comparación que ya hemos visto en el ANOVA de un factor. También se podría haber establecido contrastes planificados o *a priori*, del mismo modo que se hizo el capítulo anterior. En nuestro ejemplo, ninguno de los efectos principales resulta ser significativo, aunque sí lo es el efecto de la interacción. Para realizar comparaciones múltiples entre las interacciones, sólo se puede realizar mediante sintaxis, y para ello hay que utilizar el cuadro de diálogo de **Opciones**.

Antes de proceder a realizar comparaciones, bien de los niveles de los factores, bien entre las celdillas de las interacciones es conveniente dibujar el denominado Gráfico de perfil. Para ello en el cuadro de diálogo de Univariante (Figura 12.1) se pulsa el botón **Gráficos** si se accede al cuadro de la Figura 12.2

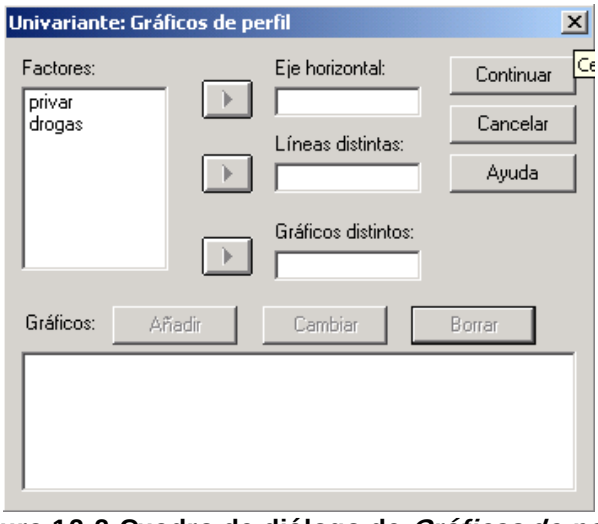


Figura 12.2 Cuadro de diálogo de *Gráficos de perfil*

En el cuadro se pasa el factor **Droga** al Eje horizontal y el factor **Privar** a Líneas distintas, luego se pulsa **Añadir**. El gráfico que se obtiene (ya retocado en el Editor de gráficos) es el que se muestra en la Figura 12.3

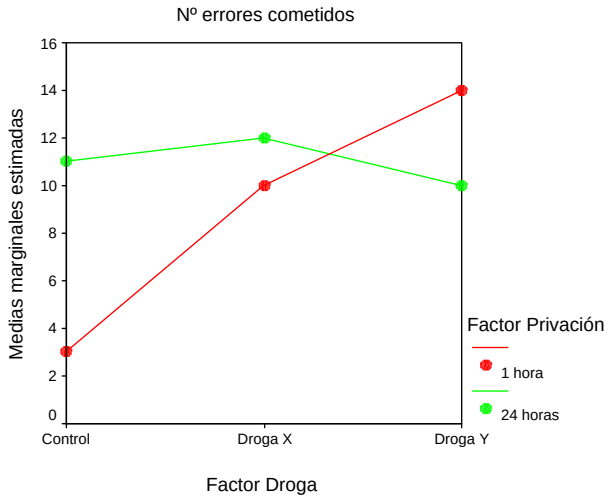


Figura 12.3 Gráfico de perfil de Droga por Privación

En el gráfico se muestras las medias del número de errores calculadas en cada uno de los seis subgrupos que resultan de combinar los niveles del factor **Droga** con los niveles del factor **Privación**. En este gráfico se ve que la diferencia entre las medias para el grupo Control de los dos grupos de Privación es amplia, o lo que es lo mismo, la diferencia en número de errores entre los dos grupos de Privación no se mantiene estable para todos los subgrupos del factor Droga.

Si cambiamos el orden de los factores, situando el factor Privar como eje horizontal y el factor Droga como líneas distintas se vería gráficamente los otros "efectos simples" del modelo.

12.3 Opciones de Univariante

Mediante las Opciones se puede obtener más información del análisis y proceder a realizar las comparaciones múltiples de la interacción, aspecto éste por el que vamos a comenzar. El cuadro de diálogo de **Opciones** es el que se muestra en la Figura 12.4.

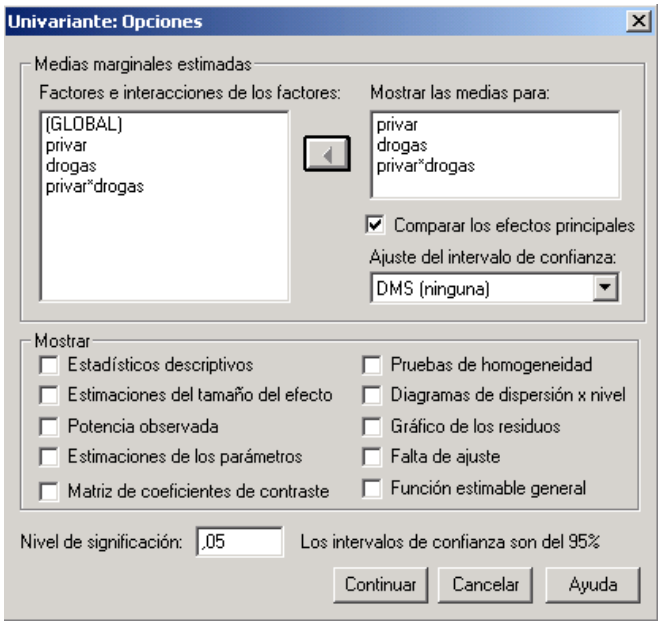


Figura 12.4 Cuadro de diálogo de Opciones de Univariante

En la lista **Factores e interacciones de los factores**, están presentes todos los factores presentes en el diseño. Una vez que se pasan los factores principales y la interacción a la lista **Mostrar medias para** se llevarían a cabo todas las comparaciones sólo para los efectos principales. Mediante el menú desplegable **Ajuste del intervalo de confianza** se puede decidir si se controla la tasa de error, es decir, la probabilidad de cometer errores tipo I en el conjunto de las comparaciones. Las dos opciones de control son la de *Bonferroni*, que controla la tasa de error multiplicando el nivel crítico concreto de cada comparación por el número de comparaciones que se están llevando a cabo entre las medias correspondientes a un mismo efecto, y la opción *Sidak* que corrige la tasa de error mediante $1-(1-p_c)^k$, donde p_c se refiere al nivel crítico de una comparación concreta y k al número de comparaciones.

Para efectuar comparaciones múltiples de la interacción es necesario **Pegar** en la ventana de sintaxis los opciones del procedimiento Univariante y realizar un pequeño añadido y ejecutar luego el procedimiento. Cuando se pega en la ventana de sintaxis, con la elección hecha en la Figura 12.4, en la ventana de sintaxis se pega lo siguiente:

```
UNIANOVA
  errores BY privar drogas
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /EMMEANS = TABLES(privar) COMPARE ADJ(LSD)
  /EMMEANS = TABLES(drogas) COMPARE ADJ(LSD)
  /EMMEANS = TABLES(privar*drogas)
  /CRITERIA = ALPHA(.05)
```

```
/DESIGN = privar drogas privar*drogas .
```

En la línea **/EMMEANS = TABLES(privar*drogas)** hay que añadir lo siguiente: **COMPARE(privar) ADJ(BONFERRONI)** de tal forma que la sintaxis quedaría de la siguiente forma:

```
UNIANOVA
  errores BY privar drogas
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /EMMEANS = TABLES(privar) COMPARE ADJ(LSD)
  /EMMEANS = TABLES(drogas) COMPARE ADJ(LSD)
  /EMMEANS = TABLES(privar*drogas) COMPARE(privar) ADJ(BONFERRONI)
  /CRITERIA = ALPHA(.05)
  /DESIGN = privar drogas privar*drogas .
```

(Si quisiéramos explorar la interacción en el otro sentido, se añadiría una nueva línea con /EMMEANS = TABLES(privar*drogas) COMPARE(drogas) ADJ(BONFERRONI))

Con esta modificación se obtiene además de las medias por cada casilla de la interacción de Privar por Droga, una tabla con las comparaciones y su significación estadística, tal como se puede ver en la Tabla 12.2.

Tabla 12.2 Comparación de la interacciones y significación estadística

Comparaciones por pares

Variable dependiente: Nº errores cometidos

Factor Droga administrada	(I) Factor Privación de comida	(J) Factor Privación de comida	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 para diferencia [a]	
						Límite inferior	Límite sup
Control	1 hora	24 horas	-8,000*	3,028	,017	-14,361	-1
	24 horas	1 hora	8,000*	3,028	,017	1,639	14
Droga X	1 hora	24 horas	-2,000	3,028	,517	-8,361	4
	24 horas	1 hora	2,000	3,028	,517	-4,361	8
Droga Y	1 hora	24 horas	4,000	3,028	,203	-2,361	10
	24 horas	1 hora	-4,000	3,028	,203	-10,361	2

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

Se ve lo que ya se detectó en el gráfico de perfil: el número de errores cometidos en el grupo Control del factor **Drogas** es significativamente menor para 1 hora de privación (3 errores en promedio) que para 24 horas de privación (11 errores en promedio). Dicha diferencia es significativa ($p = 0,017 < 0,05$).

Las demás opciones del cuadro de diálogo las comentamos de forma somera y posteriormente veremos como se visualizan los resultados en las tablas.

- **Estadísticos Descriptivos**. Se muestra la Media, desviación típica y tamaño de cada nivel y combinación de niveles. (ver Tabla 12.3).

Tabla 12.3 Estadísticos descriptivos de Univariante

Estadísticos descriptivos

Variable dependiente: Nº errores cometidos

Factor Privación	Factor Droga	Media	Desv. típ.	N
1 hora	Control	3,00	3,16	4
	Droga X	10,00	4,76	4
	Droga Y	14,00	3,92	4
	Total	9,00	5,97	12
24 horas	Control	11,00	3,92	4
	Droga X	12,00	5,48	4
	Droga Y	10,00	4,08	4
	Total	11,00	4,20	12
Total	Control	7,00	5,40	8
	Droga X	11,00	4,87	8
	Droga Y	12,00	4,28	8
	Total	10,00	5,15	24

- **Estimaciones del tamaño del efecto**. Estimaciones del grado en que cada factor o combinación está afectando a la variable dependiente. Se muestra el estadístico *eta cuadrado parcial* que se obtiene para un efecto concreto, E, de la siguiente forma: $\frac{F_E \times gl_E}{F_E \times gl_E \times gl_{ERROR}}$. Se divide pues el producto del estadístico F de ese efecto por sus grados de libertad, entre ese mismo producto y los grados de libertad del error. Este estadístico se interpreta como la proporción de varianza explicada por cada efecto. En la tabla que contiene la estimación del tamaño de los efectos (Tabla 12.4), el numerador de este estadístico se muestra, para cada efecto, en la columna Parámetro de no centralidad.

Tabla 12.4. Tabla del ANOVA con estimaciones del tamaño del efecto y Potencia

Variable dependiente: Nº errores cometidos

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta cuadrado	Parámetro de no centralidad	Potencia[a] observada
Modelo corregido	280,000[b]	5	56,000	3,055	,036	,459	15,273	,742
Intersección	2400,000	1	2400,000	130,909	,000	,879	130,909	1,000
PRIVAR	24,000	1	24,000	1,309	,268	,068	1,309	,192
DROGAS	112,000	2	56,000	3,055	,072	,253	6,109	,517
PRIVAR * DROGAS	144,000	2	72,000	3,927	,038	,304	7,855	,630
Error	330,000	18	18,333					
Total	3010,000	24						
Total corregida	610,000	23						

a. Calculado con alfa = ,05

b. R cuadrado = ,459 (R cuadrado corregida = ,309)

- **Potencia observada**. Estima la potencia asociada al contraste de cada efecto. La potencia de un contraste se refiere a su capacidad para detectar una diferencia poblacional como la diferencia muestral observada. Para el cálculo de la potencia se utiliza, por defecto, un nivel de significación de 0,05, pero puede cambiarse (Tabla 12.4).
- **Estimación de los parámetros**. Son los parámetros de los modelos ANOVA a partir de los cuales se obtiene las medias estimadas por el modelo, para cada nivel de los factores y las combinaciones entre niveles (Tabla 12.5). Estas estimaciones se obtienen combinando los parámetros involucrados en la obtención de cada media. Así, por ejemplo, para la estimación de la media del grupo Control con 1 hora de privación será: Intersección + [PRIVAR = 1] + [DROGAS = 1] + [PRIVAR = 1 * DROGAS = 1] resultando 3 que es el valor que se muestra en Tabla 12.3 (3 = 10 + 4 + 1 +(-12)). Como las estimaciones de los parámetros de cada efecto tienen suma cero la tabla no recoge los valores que son redundantes

Tabla 12.5. Estimaciones de los parámetros del modelo

Variable dependiente: Nº errores cometidos

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%.		Eta cuadrado
					Límite inferior	Límite superior	
Intersección	10,000	2,141	4,671	,000	5,502	14,498	,548
[PRIVAR=1]	4,000	3,028	1,321	,203	-2,361	10,361	,088
[PRIVAR=2]	0[b]	,	,	,	,	,	,
[DROGAS=1]	1,000	3,028	,330	,745	-5,361	7,361	,006
[DROGAS=2]	2,000	3,028	,661	,517	-4,361	8,361	,024
[DROGAS=3]	0[b]	,	,	,	,	,	,
[PRIVAR=1] * [DROGAS=1]	-12,000	4,282	-2,803	,012	-20,996	-3,004	,304
[PRIVAR=1] * [DROGAS=2]	-6,000	4,282	-1,401	,178	-14,996	2,996	,098
[PRIVAR=1] * [DROGAS=3]	0[b]	,	,	,	,	,	,
[PRIVAR=2] * [DROGAS=1]	0[b]	,	,	,	,	,	,
[PRIVAR=2] * [DROGAS=2]	0[b]	,	,	,	,	,	,
[PRIVAR=2] * [DROGAS=3]	0[b]	,	,	,	,	,	,

b. Al parámetro se le ha asignado el valor cero porque es redundante.

- **Matriz de coeficientes de contraste**. Recoge la matriz L con los coeficientes asociados a cada efecto, que son los coeficientes que definen el conjunto de hipótesis del modelo.
- **Pruebas de homogeneidad**. Muestra le estadístico de Levene de contraste de la igualdad de la varianza de los errores de las poblaciones definidas por la combinación de factores. En el caso del ejemplo (Tabla 12.6) el nivel crítico (p = 0,446 > 0,05) permite no rechazar la hipótesis de igualdad.

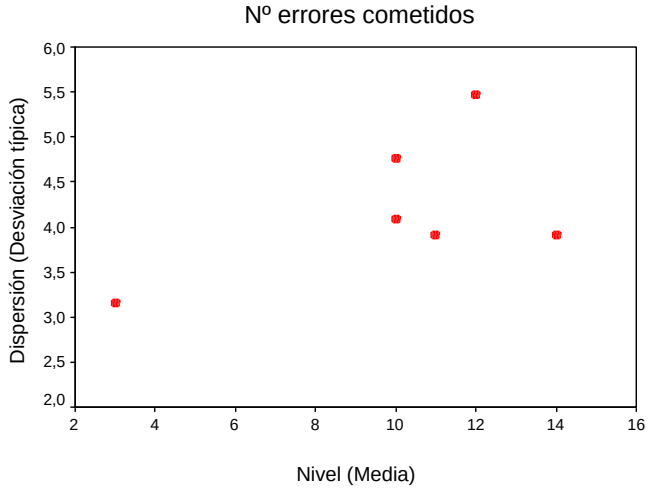
Tabla 12.6 Estadístico de *Levene* sobre igualdad de varianzas error

Variable dependiente: Nº errores cometidos

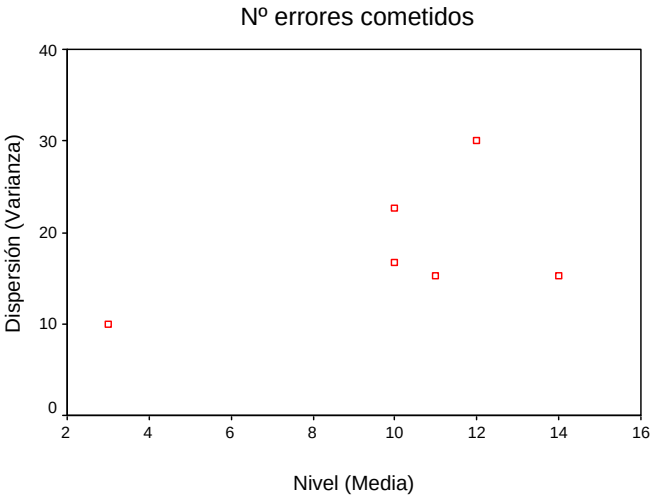
F	gl1	gl2	Significación
1,000	5	18	,446

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

- **Diagramas de dispersión por nivel**. La información gráfica de la igualdad de varianzas se puede ver gráficamente, y además se puede detectar si hay relación entre el valor de las medias y el de las varianzas. SPSS genera dos gráficos de dispersión: el de desviaciones típicas por medias, y el de varianzas por medias (Figuras 12.5 arriba y abajo).



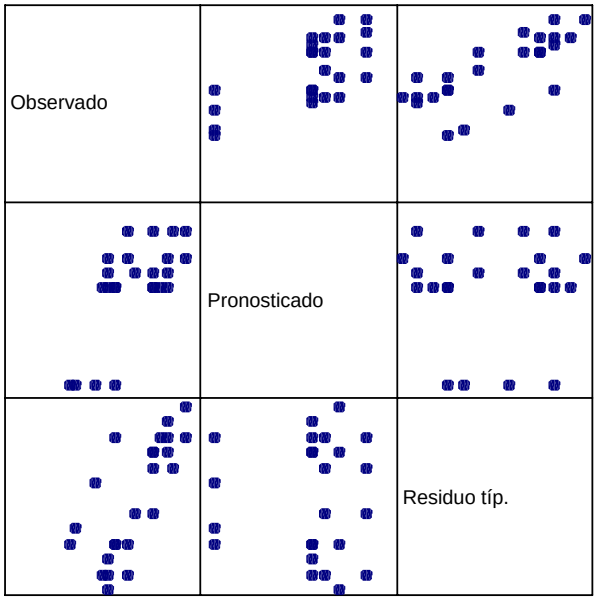
Grupos: Factor Privación de comida * Factor Droga administrada



Grupos: Factor Privación de comida * Factor Droga administrada

Figuras 12.5 Diagrama de dispersión típica por media (arriba) y diagrama de varianza por media (abajo)

- **Gráfico de residuos**. Un residuo es la diferencia entre el valor observado y el valor pronosticado por el modelo. Uno de los supuestos de los modelos ANOVA es que los residuos son independientes (no correlacionan entre sí) y se distribuyen normalmente, y además sus varianzas por nivel deben ser homogéneas, tal como hemos visto en el punto anterior. SPSS genera una matriz de gráficos en los que se relacionan los valores observados, los valores pronosticados, y los residuos tipificados. Estos gráficos no deben mostrar ninguna pauta de variación sistemática. Cuando el modelo se ajusta bien a los datos, la relación entre los valores observados y los pronosticados debe ser lineal. En la Figura 12.6 se muestra el gráfico de residuos para nuestro ejemplo.



Modelo: Intersección + PRIVAR + DROGAS + PRIVAR*DROGAS

Figura 12.6 Gráfico de residuos

12.4 Análisis de covarianza

El análisis de covarianza (ANCOVA) es una técnica de control estadístico que permite eliminar de la variable dependiente el influjo de variables no incluidas en el diseño como factores y no han sido sometidas a control experimental. La técnica del ANCOVA consiste en efectuar una regresión de la VD sobre la/s variable/s extraña/s o concomitante/s. Los errores de esta regresión serán aquella parte de la VD no explicada por las variables extrañas y por tanto libre de su influjo. Para realizar el análisis de covarianza vamos a partir del siguiente estudio:

> *Estudio sobre actitudes de escolares de secundaria. 30 alumnos (mitad mujeres: Grupo A1; mitad hombres: Grupo A2. A cada grupo se les administra dos forma paralelas de actitud hacia las cuestiones científicas, de modo que una puntuación alta refleja una alta actitud. Respecto de los tratamientos (Factor B) el Grupo 1 visionó películas sobre reportajes de ciencia; El Grupo 2 efectuó lecturas de libros científicos recomendados por los profesores; y al Grupo 3 se les mostró*

aspectos de la vida de los científicos. Las puntuaciones de pretest (X) sirvieron como covariable de las del postest (Y) que es la Variable Dependiente (VD). Se trata pues determinar si hay efectos de los factores principales y efectos de interacción una vez se ha eliminado el influjo de X sobre Y. Los datos son los siguientes:

A1						A2					
B1		B2		B3		B1		B2		B3	
Y	X	Y	X	Y	X	Y	X	Y	X	Y	X
34	18	17	21	23	16	36	16	23	30	30	31
33	15	21	20	26	17	37	19	17	34	36	32
31	15	18	19	23	18	32	15	14	29	35	31
34	19	24	18	26	21	33	19	12	32	35	30
30	15	17	21	23	20	34	20	14	27	30	27

Para realizar el ANCOVA, además de pasar las VD y los factores a sus listas correspondientes, la covariable se pasa a la lista de covariables (ver Figura 12.1).

El objetivo del análisis de covarianza sigue siendo comprobar los efectos principales de cada factor y la interacción entre ellos, pero también se puede evaluar el efecto de las covariables introducidas en el modelo, ya que el procedimiento **Univariante** permite su contraste. El estadístico F para la covariable contrasta la hipótesis de que la pendiente o coeficiente de regresión vale cero en la población.

Pueden plantearse dos situaciones: primera, si la regresión no es significativa (la pendiente vale 0), quiere decirse que la covariable no esta relacionada con la VD y por tanto puede excluirse del modelo; en este caso los resultados del ANCOVA serán similares a los del ANOVA; la segunda situación es que la pendiente sí resulte significativa. En este caso, pueden producirse dos situaciones más: la primera es que el resultado del ANOVA y del ANCOVA sean similares, lo que indica que a pesar de la relación entre la/s covariable/s y la VD y de que el influjo haya sido extraído, el efecto de las variables independientes sobre la VD no se altera por la presencia de la covariable. La segunda situación es que el resultado del ANOVA y del ANCOVA sea distinto, y esto, a su vez, puede pasar por dos motivos: un efecto significativo del ANOVA pasa a no serlo con el ANCOVA, o que un efecto no significativo con el ANOVA pasa a serlo con el ANCOVA. En el primer caso, el efecto significativo se debe al influjo de la covariable sobre la VD y no a la VD, y en el segundo caso se puede interpretar que la variable independiente, aun no estando relacionada con la VD globalmente considerada, sí lo está con la parte de la VD no explicada por la/s covariable/s.

El resumen es que siempre es conveniente realizar sobre los datos un ANOVA y tomarlo como punto de referencia para el ANCOVA. En la Tabla 12.7 se ve el resultado del ANOVA y del ANCOVA para los datos propuestos.

Tabla 12.7 Tablas resumen del ANOVA y del ANCOVA

TABLA DEL ANOVA

Variable dependiente: VD

Fuente	Suma de cuadrados tipo I	gl	Media cuadrática	F	Significación
Modelo corregido	1540,000[a]	5	308,000	39,487	,000
Intersección	21226,800	1	21226,800	2721,385	,000
FACTORA	48,133	1	48,133	6,171	,020
FACTORB	1298,600	2	649,300	83,244	,000
FACTORA * FACTORB	193,267	2	96,633	12,389	,000
Error	187,200	24	7,800		
Total	22954,000	30			
Total corregida	1727,200	29			

a. R cuadrado = ,892 (R cuadrado corregida = ,869)

TABLA DEL ANCOVA

Variable dependiente: VD

Fuente	Suma de cuadrados tipo I	gl	Media cuadrática	F	Significación
Modelo corregido	1545,837[a]	6	257,639	32,673	,000
Intersección	21226,800	1	21226,800	2691,927	,000
COVARIABLE	111,475	1	111,475	14,137	,001
FACTORA	327,309	1	327,309	41,508	,000
FACTORB	919,863	2	459,932	58,327	,000
FACTORA * FACTORB	187,190	2	93,595	11,870	,000
Error	181,363	23	7,885		
Total	22954,000	30			
Total corregida	1727,200	29			

a. R cuadrado = ,895 (R cuadrado corregida = ,868)

Antes de introducir la covariable en el modelo, los tres efectos (los dos principales y la interacción) son significativos. Después de controlar mediante la covariable, se siguen manteniendo significativos todos los efectos a pesar de que la covariable está relacionada con la VD. Es decir la relación entre la covariable y la VD no afecta al influjo de las dos variables independientes sobre la VD.

12.5 Modelos personalizados.

El procedimiento *Univariante* permite ajustar otros modelos distintos al completamente aleatorizado: modelo sin interacción, modelo con bloques aleatorios, modelos jerárquicos, etc. Algunos de estos modelos se consiguen a través del menú y otros es preciso incluir alguna modificación en la sintaxis del procedimiento. Los modelos se especifican en el cuadro de diálogo que se muestra en la Figura 12.7.

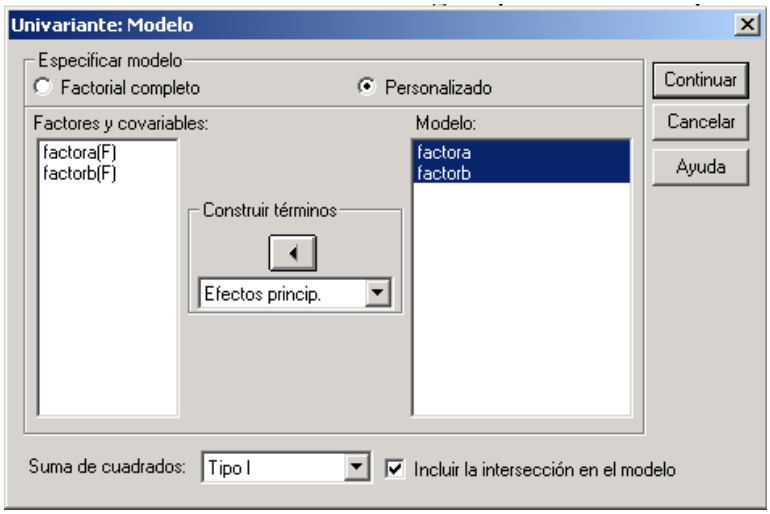


Figura 12.7 Cuadro de diálogo para especificar el Modelo

Para los modelos completamente aleatorizados sirve la opción por defecto **Factorial completo**. Para definir otros modelos es preciso marcar la opción **Personalizado**. Para cualquier modelo, sea factorial completo o personalizado, es posible elegir entre cuatro métodos distintos de suma de cuadrados. La suma de cuadrados Tipo III es la que SPSS calcula por defecto.

12.5.1 Tipos de Sumas de cuadrados

- **Tipo I**. Este método también se conoce como descomposición jerárquica del método de suma de cuadrados. Cada término se corrige sólo respecto al término que le precede en el modelo. El método Tipo I para la obtención de sumas de cuadrados se utiliza normalmente para:
  - Un modelo ANOVA equilibrado en el que se especifica cualquier efecto principal antes de cualquier efecto de interacción de primer orden, cualquier efecto de interacción de primer orden se especifica antes de cualquier efecto de interacción de segundo orden, y así sucesivamente
  - Un modelo de regresión polinómica en el que se especifica cualquier término de orden inferior antes que cualquier término de orden superior.
  - Un modelo puramente anidado en el que el primer efecto especificado está anidado dentro del segundo efecto especificado, el segundo efecto especificado está anidado dentro del tercero, y así sucesivamente. Esta forma de anidación solamente puede especificarse utilizando la sintaxis.
- **Tipo II.** Este método calcula cada suma de cuadrados del modelo considerando sólo los efectos pertinentes. Un efecto pertinente es un efecto que no está contenido en el efecto que se está examinando. El método Tipo II para la obtención de sumas de cuadrados se utiliza normalmente para:
  - Un modelo ANOVA equilibrado.
  - Cualquier modelo que sólo tenga efectos de factor principal.

- Cualquier modelo de regresión.
- Un diseño puramente anidado. Esta forma de anidación puede especificarse utilizando la sintaxis.

♦ **Tipo III.** Es el método por defecto. Este método calcula las sumas de cuadrados de un efecto del diseño como las sumas de cuadrados corregidas respecto a cualquier otro efecto que no lo contenga y ortogonales para cualquier efecto (si existe) que lo contenga. Las sumas de cuadrados de Tipo III tienen una gran ventaja por ser invariables respecto a la frecuencia de casillas, siempre que la forma general de estimabilidad permanezca constante. Así, este tipo de sumas de cuadrados se suele considerar de gran utilidad para un modelo no equilibrado sin casillas perdidas. En un diseño factorial sin casillas perdidas, este método equivale a la técnica de cuadrados ponderados de medias de **Yates**. El método Tipo III para la obtención de sumas de cuadrados se utiliza normalmente para:

- Cualquiera de los modelos que aparecen en Tipo I y Tipo II.
- Cualquier modelo equilibrado o desequilibrado sin casillas vacías.

♦ **Tipo IV.** Este método está diseñado para una situación en la que faltan casillas. Para cualquier efecto F en el diseño, si F no está contenida en cualquier otro efecto, entonces Tipo IV = Tipo III = Tipo II. Cuando F está contenida en otros efectos, el Tipo IV distribuye equitativamente los contrastes que se realizan entre los parámetros en F a todos los efectos de nivel superior. El método Tipo IV para la obtención de sumas de cuadrados se utiliza normalmente para:

- Cualquiera de los modelos que aparecen en Tipo I y Tipo II.
- Cualquier modelo equilibrado o no equilibrado con casillas vacías.

12.5.2 Modelos con bloques aleatorios

Para construir un modelo con un factor aleatorizado en bloques definidos por otro factor, hay que recordar que en este tipo de modelo el factor no interactúa con los bloques, y por tanto es un modelo sólo con efectos principales. Para definir un modelo con bloques aleatorios se siguen los siguientes pasos:

- Selección de las variables-factor y de las variables-bloque y se pasan a la lista Factores fijos.
- Pulsar botón Modelo y marcar opción Personalizado.
- Seleccionar de las variables-factor y variables-bloque en la lista **Factores** y **covariables**.
- Seleccionar, en lista de **Factores** y **covariables**, sólo las variables-factor.
- Seleccionar Todas de 2 en el menú desplegable Construir términos y pulsar la flecha para pasar a la lista Modelo todas las combinaciones entre cada dos variables-factor (se dejan fueran todas las combinaciones entre factor y bloque y entre variables-bloque entre sí).

- Si se trata de un modelo de más de dos factores, se debería seleccionar en Construir términos la opción Todas de 3, y así sucesivamente.

12.5.3 Modelos jerárquicos o anidados

En este tipo de diseños uno de los factores está anidado en el otro factor, lo que significa que los niveles de uno de los factores son distintos en cada nivel del otro factor. Por esta razón no es posible evaluar la interacción, pero sí los efectos principales. Los pasos son los siguientes:

- Seleccionar las variables independientes pasarlas a la lista **Factores fijos**.
- Pulsar botón **Modelo** y marcar opción Personalizado.
- Seleccionar la variable que define el factor no anidado en la lista Factores y covariables.
- Seleccionar Efectos principales en el menú Construir términos y trasladar la variable a la lista Modelo.
- Pegar las selecciones hechas en una ventana de sintaxis y editar. En el procedimiento, en el submandato DESIGN estará el factor no anidado. Añadir a esta línea el nombre del factor anidado y, entre paréntesis, el nombre del factor no anidado. Es decir DESIGN-factor no anidado - factor anidado (factor no anidado). Por ejemplo, si en el ejemplo de drogas y privación, privación fuera el factor no anidado y drogas el anidado, la sintaxis sería:

"DESIGN privar drogas(privar)"

12.5.4 Homogeneidad de las pendientes de regresión

En el ANCOVA, en cada nivel de la variable independiente o combinación de variables independientes, hay una ecuación de regresión que relaciona la variable dependiente y la covariable. Uno de los supuestos del ANCOVA es la homogeneidad de las pendientes de estas rectas de regresión, o lo que es lo mismo la relación entre VD y covariable es la misma para cada nivel o combinación de niveles. Con SPSS se puede contrastar este supuesto, siguiendo los siguientes pasos:

- Seleccionar la variable dependiente y pasarla al cuadro correspondiente.
- Seleccionar la variable independiente y pasarla a la lista **Factores fijos**.
- Seleccionar la covariable y trasladarla a la lista **Covariables**
- Pulsar **Modelo** y seleccionar **Personalizado**.
- Seleccionar la variable independiente y la covariable de la lista **Factores y covariab**les.
- Seleccionar **Efectos principales** en el menú **Construir términos** y pasar las variables a la lista **Modelo**.
- Seleccionar de nuevo la variable independiente y la covariable de la lista **Factores y covariables**.
- Seleccionar **Interacción** en el menú **Construir términos** y pasar las variables a la lista Modelo.

El estadístico F referido a la interacción señalará si la hipótesis de igualdad de pendientes es o no aceptada. Si se asume la hipótesis se procede al ANCOVA, ya que este modelo está basado en la estimación de una única recta de regresión para todos los niveles o combinaciones de niveles. Si, por el contrario, se rechaza la hipótesis de homogeneidad de pendientes, en el modelo de análisis de covarianza se deben incorporar estimaciones separadas para cada pendiente. Para ello el proceso es el mismo que los 8 pasos anteriores, pero después del octavo paso:

- Pulsar **Modelo** y seleccionar **Personalizado**.
- Seleccionar la variable independiente de la lista **Factores y covariab**les.
- Seleccionar **Efectos principales** en el menú **Construir términos** y pasar la variable a la lista **Modelo**.
- Seleccionar de nuevo la variable independiente y la covariable de la lista **Factores y covariables**.
- Seleccionar **Interacción** en el menú **Construir términos** y pasar las variables a la lista Modelo.
- Desmarcar la opción Incluir la intersección en el modelo.

12.6 Contrastes personalizados

Estos contrastes permiten comparaciones más amplias que los contraste *post hoc*, como ya vimos en el procedimiento **Anova**. Para realizar estos contrastes se pulsa en botón Contrastes del cuadro de diálogo de **Univariante** y se muestra el cuadro de la Figura 12.8.

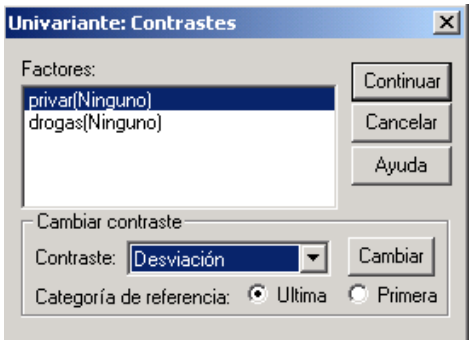


Figura 12.8 Cuadro de diálogo de *Contrastes*

La lista Factores contiene un listado de los factores seleccionados. Por defecto, los factores no tienen asignado ningún tipo de contraste. Para asignarle un tipo, se marca el factor y se seleccionar del menú desplegable **Contraste** el contraste deseado y se pulsa el botón **Cambiar**. Con todas las opciones se obtienen k – 1 comparaciones entre los k niveles de un factor, pero cada contraste permite un tipo particular de comparaciones:

- **Desviación**. Todas las categorías de un factor excepto una (por defecto, la última) se comparan con la media total. Se puede elegir también la primera categoría como Categoría de referencia. Para otras alternativas hay que recurrir a la sintaxis. En la línea CONTRASTE, después del tipo de contraste se pone, entre paréntesis, la categoría que se quiere utilizar como referencia.

- **Simple**. Cada categoría, excepto la última, se compara con la última categoría. También se puede elegir la primera categoría, o cualquier otra, siguiendo el procedimiento sintáctico explicado en el anterior contraste.
- **Diferencia**. Cada categoría, menos la primera, se compara con la media de las categorías anteriores. En los diseños equilibrados, las k – 1 comparaciones son ortogonales.
- **Helmert**. Cada categoría, menos la última, se compara con la media de las categorías posteriores. En los diseños equilibrados, las k – 1 comparaciones son ortogonales.
- **Repetido**. Cada categoría, excepto la primera, se compara con la categoría anterior.
- **Polinómico**. Se compara la tendencia. La primera es la lineal, la segunda la cuadrática, etc. En un diseño equilibrado los contrastes son ortogonales.
- **Especial**. Sólo está disponible mediante sintaxis. Es la forma de hacer contrastes más complejos. Para ello en la línea CONTRAST se añade la palabra SPECIAL y entre paréntesis el contraste que queremos efectuar. Si por ejemplo, queremos contrastar, en el factor drogas, el grupo control con el de droga X y el grupo control con el de droga Y añadiríamos "CONTRAST (droga) = SPECIAL(1 –1 0 1 0 –1).

13. El Modelo Lineal General.

Análisis de varianza con medidas repetidas.

13.1 Introducción

En el capítulo anterior hemos visto, dentro del modelo lineal general, lo que en el ámbito del diseño se denominan diseños factoriales completamente aleatorizados, en donde se asigna aleatoriamente a los sujetos a cada una de las condiciones experimentales que se deriven del número de factores que concurren en el diseño. En este caso las fuentes de variabilidad del análisis representan las diferencias entre los sujetos sometidos a las distintas condiciones. En contraste con esto, hay otro tipo de diseños, en los que los sujetos pueden servir en todas o en algunas de las condiciones o tratamientos. En estas circunstancias, parte de la variabilidad extraída en el análisis será el reflejo de la variabilidad de cada sujeto, y esta es la razón por la cual a este tipo de diseños se les denomina *diseños intra-sujetos* o diseños de medidas repetidas.

En las ciencias del comportamiento, una alta proporción de los diseños que se llevan a cabo son de este tipo por la ventajas que suponen. Son diseños que permiten estudiar cambios en conductas tales como aprendizaje, entrenamiento, recuerdo, cambio de actitudes, etc., y que son particularmente sensibles y eficientes, en el sentido de que son más económicos en cuanto al número de sujetos, en comparación con los diseños *entre-sujetos*.

No obstante, también presentan ciertos inconvenientes que hay que señalar: por un lado, está el **efecto de la práctica**, debido al cual los sujetos pueden mostrar una mejora general durante el transcurso de la prueba, o bien pueden mostrar fatiga y empeorar la ejecución. De este efecto "perverso" sólo estaría inmune el primer ensayo, y es por ello por lo que nunca se usa el mismo orden al administrar los tratamientos a los sujetos; una técnica que evita este efecto es el contrabalanceo (por ejemplo, los diseños de cuadrado latino).

El otro inconveniente es el denominado **efecto de arrastre**, el cual puede no afectar de la misma manera a todas las condiciones, y que no se puede contrarrestar mediante el contrabalanceo. Pensemos, por ejemplo, en el efecto no disipado que puede tener una droga concreta sobre los demás tratamientos, o el efecto del alguna instrucción dada al principio que es recordada por el sujeto posteriormente e influye en la ejecución en los tratamientos posteriores. Ante esto la única alternativa es alargar el tiempo entre tratamientos de modo que se disipe completamente el efecto del tratamiento anterior.

En este capítulo, vamos a tratar los tres tipos básicos de diseños de medidas repetidas. En primer lugar, el diseño de un factor con medidas repetidas (1 Factor MR); en segundo lugar, el diseño de dos factores, uno completamente aleatorizado y el otro con medidas repetidas (2 Factores, uno MR y otro CA); y por último, el diseño de dos factores con medidas repetidas en ambos (2 Factores MR). En el Cuadro 13.1 se realiza una comparación de los tres tipos de diseños y su notación.

Cuadro 13.1 Comparación de los diseños intra-sujetos

Diseño A x S
Un factor intra-sujetos

	a_1	a_2	a_3
s_1			
s_2			
s_3			

Diseño A x (B x S)
Un factor entre-sujetos y otro intra-sujetos
(Diseño mixto o diseño split-plot)

	a_1 b_1	a_1 b_2	a_1 b_3
s_1			
s_2			
s_3			

	a_2 b_1	a_2 b_2	a_2 b_3
s_4			
s_5			
s_6			

Diseño (A x B x S)
Dos factores intra-sujetos

	a_1 b_1	a_1 b_2	a_1 b_3	a_2 b_1	a_2 b_2	a_2 b_3
s_1						
s_2						
s_3						

La letras A y B designan los factores y S designa a los sujetos

13.2 Diseño de un factor intra-sujetos

Es el caso más simple: un grupo de sujetos pasa por todas las condiciones o niveles de un único factor. Para llevar a cabo el análisis vamos a utilizar un conjunto de datos basados en un experimento de atención, en el cual, a cada sujeto, se le dan cinco páginas impresas, asignadas en sentido aleatorio. Cada página tienen un nivel de dificultad de lectura ascendente (menos difícil el nivel 1 y más difícil el nivel 4) , y cada una contiene el mismo número de errores tipográficos. Los sujetos tenían que localizar los errores y la VD es el número de errores identificados correctamente. Los datos son los siguientes:

	Nivel dificultad lectura			
Sujetos	1	2	3	4
1	14	12	7	6
2	15	10	9	9
3	16	8	11	9
4	13	11	8	9
5	16	12	7	12

6	**16**	**10**	**8**	**11**
7	**14**	**13**	**12**	**10**
8	**12**	**8**	**11**	**7**
9	**11**	**8**	**8**	**10**

Para introducir datos con medidas repetidas (MR) en el editor de datos de SPSS, cada una de las condiciones o niveles del factor se corresponde con una variable, de tal modo que el archivo constará de tantos casos como sujetos y tantas variables como condiciones experimentales o niveles del factor. El lector recordará, que en el caso de un factor completamente aleatorizado (CA), el factor se corresponde con una sola variable y los valores de la variable son los niveles del factor. Para estos datos, la disposición en el editor de SPSS sería la que se muestra en la Figura 13.1

	a1	a2	a3	a4
1	14	12	7	6
2	15	10	9	9
3	16	8	11	9
4	13	11	8	9
5	16	12	7	12
6	16	10	8	11
7	14	13	12	10
8	12	8	11	7
9	11	8	8	10

Figura 13.1 Datos para un diseño de un factor con medidas repetidas

Para llevar a cabo el análisis hay que seguir la siguiente secuencia:

Analizar → Modelo Lineal General → Medidas repetidas...

y se accede al cuadro de diálogo de la Figura 13.2

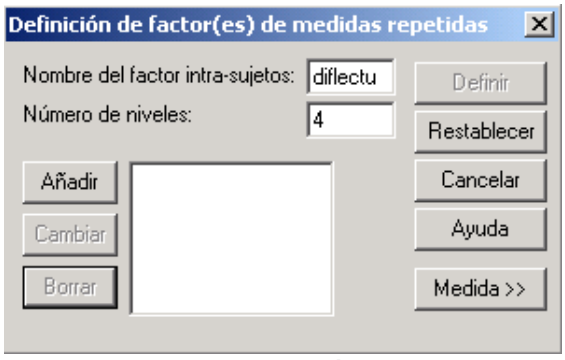


Figura 13.2 Cuadro de diálogo para definir factores de medidas repetidas

En este primer cuadro se definen el factor o factores *intra-sujetos* y los niveles. Después de dar nombre al factor y especificar el número de niveles, se pulsa el botón **Añadir** y se traslada a la lista central. Cuando que hay un factor trasladado a la lista se activa el botón **Definir**, que al pulsar, se accede al cuadro de diálogo de la Figura 13.3

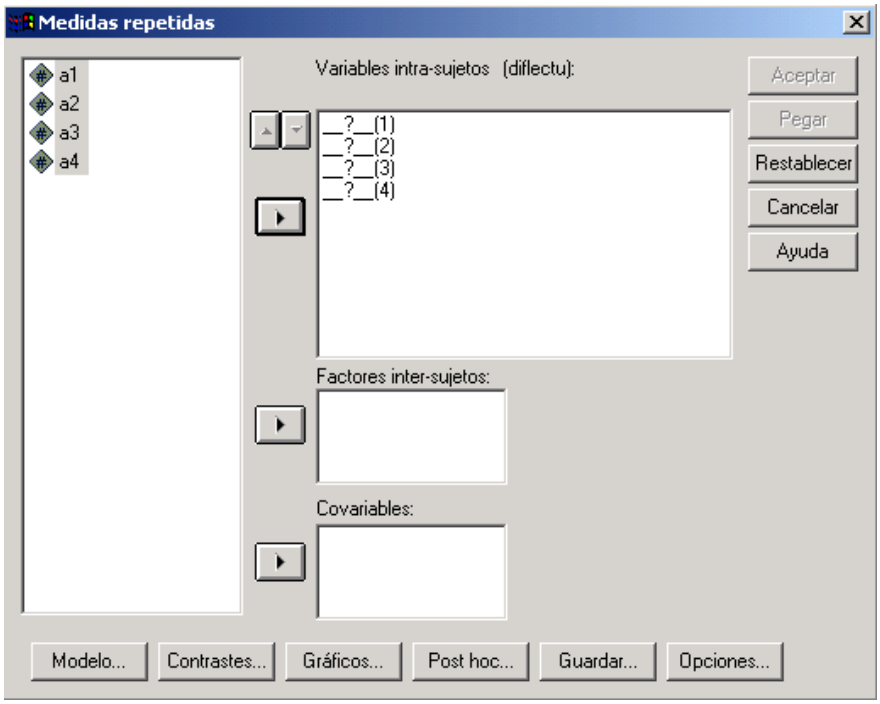


Figura 13.3 Cuadro de diálogo de Medidas repetidas

A la lista **Variables intra-sujetos** se trasladan los nombres de las variables que definen el factor *intra-sujetos*, que son las que hemos definido en el cuadro anterior. Una vez pasadas las variables, el orden de estas se puede modificar con los botones .

En la lista **Factores inter-sujetos** se incluirán los factores que hubiere de este tipo. Y, del mismo modo, para el caso en que el diseño contenga alguna variable concomitante, que se incluiría en la lista de **Covariables**.

Una vez realizada y aceptada la selección del factor intra-sujetos en el Visor de resultados se ofrecen, por defecto, varias tablas que se muestran de manera conjunta en las tablas de la Tabla 13.1.

Tabla 13.1. Resultados por defecto de Medidas repetidas

Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación
DIFLECTU	Traza de Pillai	,898	17,582	3,000	6,000	,002
	Lambda de Wilks	,102	17,582	3,000	6,000	,002
	Traza de Hotelling	8,791	17,582	3,000	6,000	,002
	Raíz mayor de Roy	8,791	17,582	3,000	6,000	,002

Prueba de esfericidad de Mauchly

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon		
					Greenhouse -Geisser	Huynh-Feldt	Límite-inferior
DIFLECTU	,819	1,344	5	,931	,885	1,000	,333

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
DIFLECTU	Esfericidad asumida	152,306	3	50,769	15,632	,000
	Greenhouse-Geisser	152,306	2,656	57,342	15,632	,000
	Huynh-Feldt	152,306	3,000	50,769	15,632	,000
	Límite-inferior	152,306	1,000	152,306	15,632	,004
Error(DIFLECTU)	Esfericidad asumida	77,944	24	3,248		
	Greenhouse-Geisser	77,944	21,249	3,668		
	Huynh-Feldt	77,944	24,000	3,248		
	Límite-inferior	77,944	8,000	9,743		

La primera tabla es la de los contrastes multivariados, que ofrece cuatro estadísticos: la *Traza de Pillai*, la *Lamdba de Wilks*, la *Traza de Hotelling* y la *Raíz mayor de Roy*. Una explicación de estos estadísticos se encuentra en Bock (1975) y Tabachink y Fidel (1983). Su interpretación es la misma que la de cualquier estadístico: si el nivel crítico es menor de 0,05 se rechaza la hipótesis de igualdad de medias de los tratamientos. En nuestro caso, efectivamente, se puede decir que el número de errores tipográficos detectados depende del nivel de dificultad de lectura de la página.

Uno de los supuestos básicos de los modelos de medidas repetidas es la igualdad de las varianzas de las diferencias entre cada dos niveles del factor. Como hay 4 condiciones tendremos 6 combinaciones dos a dos, es decir, tendremos 6 nuevas variables cuyas varianzas se supone que serán iguales. Este supuesto se denomina de *circularidad de la matriz de varianzas-covarianzas* -hay varios textos que lo explican; recomendamos el de Kirk (1995), Winner, Brown y Michels (1991) y Keppel (1991). Para contrastar el supuesto, el procedimiento **Medidas repetidas** aporta la *prueba de esfericidad W de Mauchly* (segunda tabla), que para los datos que se analizan nos lleva a aceptar dicha hipótesis.

En el caso de que el estadístico *W* lleve al rechazo de la hipótesis, se ofrecen dos soluciones alternativas. La primera es basar la decisión en los contrastes multivariados, que no están afectados por el incumplimiento de dicho supuesto. La segunda es utilizar el estadístico *F* univariado aplicando un factor de corrección denominado *Épsilon* (ver tabla segunda de la Tabla 13.1), el cual expresa el grado en que la matriz de varianzas-covarianzas se aleja de la esfericidad. Son dos las estimaciones de épsilon: la de *Greenhouse-Geisser* (1958) y la *Huynh-Feldt* (1976). El tercer valor ofrecido, *Límite inferior*, es el valor más extremo que alcanzaría *épsilon* en el caso de un incumplimiento máximo de la esfericidad. Para su utilización, se multiplica el valor de *épsilon* por los grados de libertad de la F (numerador y denominador). Obviamente, el valor observado de la *F* resultante es idéntico en todos los casos (se multiplica y divide por el mismo factor *épsilon*), pero al disminuir los grados de libertad también disminuye el nivel crítico, es decir, en el caso de que se incumpla el supuesto de circularidad la *F* observada tiene que ser mayor que cuando se cumple dicho supuesto para que se pueda rechazar la hipótesis de igualdad de medias entre los tratamientos.

13.2.1 Modelo y contrastes

Además de las opciones por defecto el procedimiento ***Medidas repetidas*** ofrece otras opciones que ya estaban presentes en el Anova factorial. La primera de ellas es el **Modelo**, pero en el caso de un factor, carece de sentido pues el único modelo posible es el que incluye el factor.

Respecto de los contrastes, el procedimiento ofrece por defecto contrastes de tipo **Polinómico**, que permite analizar la tendencia de los datos (lineal, cuadrática, cúbica, etc.). Si no se modifica esta opción por defecto la tabla con los contrastes **Polinómicos** es la que se ve en la Tabla 13.2 y se denomina **Prueba de los contrastes intra-sujetos**.

Tabla 13.2 Contrastes Polinómicos intra-sujetos

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	DIFLECTU	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
DIFLECTU	Lineal	113,606	1	113,606	47,974	,000
	Cuadrático	38,028	1	38,028	11,385	,010
	Cúbico	,672	1	,672	,167	,694
Error(DIFLECTU)	Lineal	18,944	8	2,368		
	Cuadrático	26,722	8	3,340		
	Cúbico	32,278	8	4,035		

Al tratarse de contrastes ortogonales, se ofrecen tantos contrastes como niveles tiene el factor, menos 1. En nuestro ejemplo, los contrastes significativos son el lineal y el cuadrático. En el caso de varios componentes significativos se interpreta el de mayor orden. (Observe el lector que la suma de cuadrados de los tres contrastes es lógicamente igual a la suma de cuadrados obtenida en la prueba de los efectos intra-sujetos: 113,606 + 38,028 + 0,672 = 152,306).

Además, y también como opción por defecto, se ofrece la prueba de los efectos inter-sujetos, que contrasta la hipótesis de que la media poblacional global vale cero, aunque en general este es un contraste que carece de sentido. El contraste se puede ver en la Tabla 13.3

Tabla 13.3 Efectos inter-sujetos

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1

Variable transformada: Promedio

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Intercept	4074,694	1	4074,694	957,188	,000
Error	34,056	8	4,257		

13.2.2 Gráficos de perfil

Estos gráficos ayudan a comprender el resultado de los contrastes Polinómicos. Para obtenerlo se pulsa en el botón **Gráficos** en el cuadro de diálogo de **Medidas repetidas** (ver Figura 13.3) y se accede al cuadro de la Figura 13.4.

Figura 13.4 Cuadro de diálogo de gráficos de perfil

Para generar el gráfico se pasa el factor al cuadro **Eje horizontal** y luego se pulsa el botón **Añadir** para pasarlo a la lista inferior. El gráfico resultante es el que se muestra en la Figura 13.5.

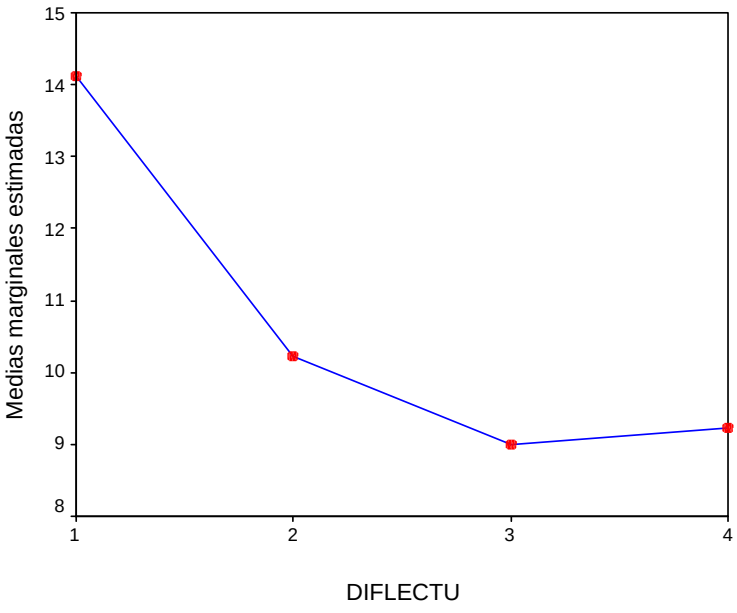


Figura 13.5. Gráfico de perfil representando el nivel de dificultad de lectura

En el gráfico se puede ver que el número de errores tipográficos detectados disminuye a medida que aumenta el nivel de dificultad de lectura de la página.

13.2.3 Opciones

En los diseños de un factor con MR al no existir grupos (como en el caso de un diseño inter-sujetos CA) no se pueden efectuar comparaciones post-hoc, pero si se

puede realizar comparaciones de los efectos principales entre pares de niveles del factor. Para ello se pulsa el botón **Opciones** del cuadro de diálogo de Medidas repetidas (Figura 13.3). y se accede al cuadro de la Figura 13.6.

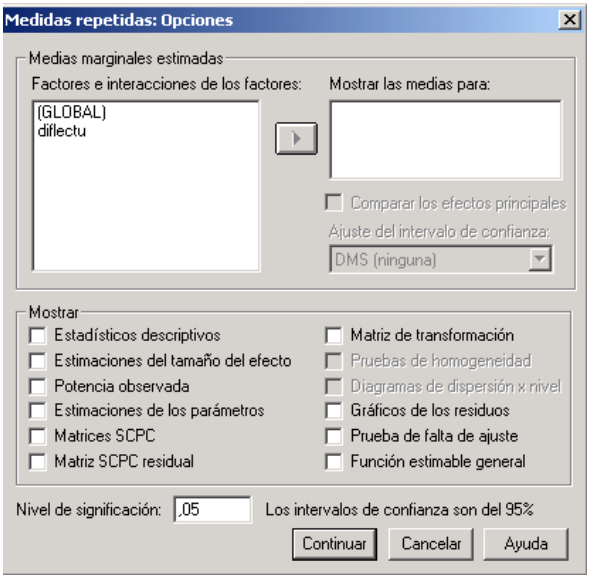


Figura 13.6. Cuadro de diálogo de opciones de MR

En este cuadro se selecciona la variable **diflectu** en la lista de factores e interacciones de los factores y se traslada a la lista **Mostrar las medias para**. Se marca la opción **Comparar los efectos principales** (que se activa cuando hay algún factor trasladado), y dentro del menú de persiana **Ajuste del intervalo**, se elige la opción **Bonferroni**. Con sólo esas elecciones las tablas que se muestran en el Visor son las de la Tabla 13.4.

Tabla 13.4 Tablas de comparación de los efectos principales

Estimaciones

Medida: MEASURE_1

DIFLECTU	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
1	14,111	,611	12,702	15,520
2	10,222	,641	8,745	11,700
3	9,000	,624	7,562	10,438
4	9,222	,619	7,796	10,649

Comparaciones por pares

Medida: MEASURE_1

(I) DIFLECTU	(J) DIFLECTU	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 % para diferencia [a]	
					Límite inferior	Límite superior
1	2	3,889	,735	,004	1,332	6,446
	3	5,111	,904	,003	1,965	8,257
	4	4,889	,676	,001	2,538	7,240
2	1	-3,889	,735	,004	-6,446	-1,332
	3	1,222	,983	1,000	-2,197	4,642
	4	1,000	,816	1,000	-1,840	3,840
3	1	-5,111	,904	,003	-8,257	-1,965
	2	-1,222	,983	1,000	-4,642	2,197
	4	-,222	,940	1,000	-3,491	3,046
4	1	-4,889	,676	,001	-7,240	-2,538
	2	-1,000	,816	1,000	-3,840	1,840
	3	,222	,940	1,000	-3,046	3,491

a. Ajuste para comparaciones múltiples: Bonferroni.

En la primera de las tablas se ofrecen las medias estimadas para cada nivel del factor MR, junto con el error típico y el intervalo de confianza, y en la segunda las comparaciones por pares entre niveles. Se observa que el nivel 1 es significativamente diferente del 2, 3 y 4, mientras que no hay diferencias entre ellos. Esto también se ve en los intervalos de confianza de la tabla de las medias estimadas: el del nivel 1 no se solapa con ninguno de los otros niveles, y sin embargo sí se solapan entre ellos.

Otra de las opciones de MR es la denominada **Matriz de transformación**, que muestra los coeficientes asignados a los niveles del factor en cada uno de los posibles contrastes polinómicos. También, como otra opción más, están las **Matrices SCPC** (matrices de suma de cuadrados y productos cruzados), y proporciona una para cada efecto inter-sujetos, para cada efecto intra-sujetos y para el término error. Para un efecto dado, dicha matriz muestra, en la diagonal, la suma de cuadrados correspondiente a ese efecto descompuesta en tantas partes como grados de libertad tiene el efecto. La descomposición se efectúa a partir de los contrastes definidos en el cuadro de diálogo de **Contrastes**. En las tablas de la Tabla 13.5 se muestran las 3 tablas que se generan cuando se marcan estas opciones. Además, siempre que se solicita la Matriz SCPC residual, se muestra la prueba de esfericidad de *Barlett*, similar a la prueba de *Mauchly* y que contrasta la hipótesis de que la matriz de varianzas-covarianzas residual es proporcional a una matriz identidad.

Tabla 13.5 Tablas con las matrices de transformación y la de SCPC

Medida: MEASURE_1

Variable dependiente	DIFLECTU		
	Lineal	Cuadrático	Cúbico
A1	-,671	,500	-,224
A2	-,224	-,500	,671
A3	,224	-,500	-,671
A4	,671	,500	,224

Matriz SCPC del efecto del factor intra-sujeto

			DIFLECTU : Columna		
		DIFLECTU : fila	Lineal	Cuadrático	Cúbico
Hipótesis	Intercept	Lineal	113,606	-65,728	8,739
		Cuadrático	-65,728	38,028	-5,056
		Cúbico	8,739	-5,056	,672
Error		Lineal	18,944	,323	-6,389
		Cuadrático	,323	26,722	6,733
		Cúbico	-6,389	6,733	32,278

Matriz SCPC residual

		A1	A2	A3	A4
Suma de cuadrados y productos cruzados	A1	26,889	8,778	-2,000	10,778
	A2	8,778	29,556	-6,000	4,556
	A3	-2,000	-6,000	28,000	-4,000
	A4	10,778	4,556	-4,000	27,556
Covarianza	A1	3,361	1,097	-,250	1,347
	A2	1,097	3,694	-,750	,569
	A3	-,250	-,750	3,500	-,500
	A4	1,347	,569	-,500	3,444
Correlación	A1	1,000	,311	-,073	,396
	A2	,311	1,000	-,209	,160
	A3	-,073	-,209	1,000	-,144
	A4	,396	,160	-,144	1,000

13.3 Modelo de dos factores, uno con medidas repetidas

Este tipo de modelos de dos factores, uno CA y otro MR, se conocen en el ámbito de los Psicología como modelo mixto, y en el ámbito de la estadística como modelo split-plot. Para realizar el análisis vamos a utilizar los siguientes datos:

Factor B (CA)		Negro					Verde					Azul			
Factor A (MR)		1	2	3	4		1	2	3	4		1	2	3	4
	1	14	12	7	6	10	13	12	8	7	19	12	9	7	8
	2	15	10	9	9	11	10	10	5	7	20	11	5	7	7
	3	16	8	11	9	12	9	8	5	8	21	8	5	4	3
	4	13	11	8	9	13	10	8	7	8	22	10	4	4	4
	5	16	12	7	12	14	12	9	5	5	23	7	6	6	5
	6	16	10	8	11	15	13	8	10	5	24	6	3	3	3
	7	14	13	12	10	16	10	7	6	4	25	8	4	4	5
	8	12	8	11	7	17	12	9	6	4	26	8	5	5	3
	9	11	8	8	10	18	14	7	4	5	27	5	3	2	1

que son una continuación de los datos que hemos utilizado para un factor MR. En este caso, el factor intra-sujetos sigue siendo la dificultad de lectura de la página, y

el factor inter-sujetos, el color de letra utilizado para escribir la página. Como VD seguimos utilizando el número de errores tipográficos correctamente detectados.

Para introducir en el editor los datos para un diseño mixto de un factor intra y otro inter, necesitamos, además de las cuatro variables del factor intra, una variable adicional que contenga los niveles del factor inter. En nuestro caso a este factor lo hemos denominado color, y los datos en el editor de datos de SPSS quedaría como se muestra en la Figura 13.7. Los niveles del factor intra los hemos denominado **diflect1** a **diflect4**.

	color	diflect1	diflect2	diflect3	diflect4
1	Negro	14	12	7	6
2	Negro	15	10	9	9
3	Negro	16	8	11	9
4	Negro	13	11	8	9
5	Negro	16	12	7	12
6	Negro	16	10	8	11
7	Negro	14	13	12	10
8	Negro	12	8	11	7
9	Negro	11	8	8	10
10	Verde	13	12	8	7
11	Verde	10	10	5	7
12	Verde	9	8	5	8
13	Verde	10	8	7	8
14	Verde	12	9	5	5
15	Verde	13	8	10	5
16	Verde	10	7	6	4
17	Verde	12	9	6	4
18	Verde	14	7	4	5
19	Azul	12	9	7	8
20	Azul	11	5	7	7
21	Azul	8	5	4	3
22	Azul	10	4	4	4
23	Azul	7	6	6	5
24	Azul	6	3	3	3
25	Azul	8	4	4	5
26	Azul	8	5	5	3
27	Azul	5	3	2	1

Figura 13.7 Datos de un ANOVA mixto: un factor MR y un factor CA

Para realizar el análisis, en el cuadro de diálogo de Medidas repetidas (ver Figura 13.3) se selecciona el factor inter-sujetos (color) y se traslada a la lista Factores inter-sujetos. En el Visor, los resultados que se muestran por defecto, son los que se muestran en las tablas de la Tabla 13.6.

Tabla 13.6. Tablas de resultados por defecto de un diseño mixto en el procedimiento Medidas repetidas

Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación
DIFLECT	Traza de Pillai	,880	54,002	3,000	22,000	,000
	Lambda de Wilks	,120	54,002	3,000	22,000	,000
	Traza de Hotelling	7,364	54,002	3,000	22,000	,000
	Raíz mayor de Roy	7,364	54,002	3,000	22,000	,000
DIFLECT * COLOR	Traza de Pillai	,303	1,371	6,000	46,000	,246
	Lambda de Wilks	,709	1,376	6,000	44,000	,246
	Traza de Hotelling	,393	1,375	6,000	42,000	,247
	Raíz mayor de Roy	,342	2,620	3,000	23,000	,075

Prueba de esfericidad de Mauchly

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Significación	Epsilon		
					Greenhouse -Geisser	Huynh-Feldt	Límite-inf
DIFLECT	,970	,687	5	,984	,981	1,000	

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
DIFLECT	Esfericidad asumida	405,731	3	135,244	59,842	,000
	Greenhouse-Geisser	405,731	2,942	137,890	59,842	,000
	Huynh-Feldt	405,731	3,000	135,244	59,842	,000
	Límite-inferior	405,731	1,000	405,731	59,842	,000
DIFLECT * COLOR	Esfericidad asumida	19,796	6	3,299	1,460	,204
	Greenhouse-Geisser	19,796	5,885	3,364	1,460	,206
	Huynh-Feldt	19,796	6,000	3,299	1,460	,204
	Límite-inferior	19,796	2,000	9,898	1,460	,252
Error(DIFLECT)	Esfericidad asumida	162,722	72	2,260		
	Greenhouse-Geisser	162,722	70,618	2,304		
	Huynh-Feldt	162,722	72,000	2,260		
	Límite-inferior	162,722	24,000	6,780		

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	DIFLECT	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
DIFLECT	Lineal	334,491	1	334,491	148,754	,000
	Cuadrático	70,083	1	70,083	33,980	,000
	Cúbico	1,157	1	1,157	,469	,500
DIFLECT * COLOR	Lineal	10,693	2	5,346	2,378	,114
	Cuadrático	3,167	2	1,583	,768	,475
	Cúbico	5,937	2	2,969	1,202	,318
Error(DIFLECT)	Lineal	53,967	24	2,249		
	Cuadrático	49,500	24	2,062		
	Cúbico	59,256	24	2,469		

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1
Variable transformada: Promedio

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Intercept	7056,750	1	7056,750	979,915	,000
COLOR	465,167	2	232,583	32,297	,000
Error	172,833	24	7,201		

La primera tabla presenta los **estadísticos multivariados**, ya comentados en el apartado anterior. La siguiente, es la **prueba de esfericidad de Mauchly**, y dado que el nivel crítico es mayor de 0,05 se acepta la hipótesis de esfericidad, por lo que se puedes basar las decisiones sobre los estadísticos univariados.

Las siguientes dos tablas, se refiere al contraste de los efectos. La primera es la prueba de los **efectos intra-sujetos** y su interacción con el factor inter-sujetos, según la cual, se rechaza la hipótesis de que las medias de errores son iguales en las cuatro condiciones de lectura ($F = 59,842$; $p < 0,05$), pero no se puede rechazar la hipótesis de que no hay interacción con el factor inter-sujetos ($F = 1,460$; $p > 0,05$). La segunda prueba es la de los efectos inter-sujetos, y según esta prueba sí se rechaza la hipótesis de los promedios de los errores son iguales en las tres condiciones de color ($F = 32,297$; $p < 0,05$).

13.3.1 Pruebas de homogeneidad de varianzas

Además de las pruebas de los efectos, se pueden obtener contrastes sobre alguno de los supuestos adicionales en este tipo de diseño, como es el de la **igualdad de matriz de varianzas-covarianzas de los niveles del factor intra-sujetos**. Para contrastar este supuesto se marca la opción **Pruebas de homogeneidad** en el cuadro de diálogo de **Opciones** (ver Figura 13.6), después de haber pasado el factor intra-sujetos a la lista **Mostrar medias para**. Con esta opción marcada, en el Visor se ofrecen dos estadísticos: el de *Box* y el de *Levene*. En las tablas de la Tabla 13.7 se ven los valores de los dos estadísticos.

Tabla 13.7 Tablas con las pruebas Box y de Levene

Prueba de Box de igualdad de matrices de varianzas-covarianzas

M de Box	26,460
F	,994
gl1	20
gl2	2067,585
Significación	,466

Contraste de Levene sobre la igualdad de las varianzas error

	F	gl1	gl2	Significación
DIFLECT1	,257	2	24	,776
DIFLECT2	,362	2	24	,700
DIFLECT3	,071	2	24	,931
DIFLECT4	,283	2	24	,756

13.3.2 Gráficos de perfil

Para obtener gráficamente los promedios combinados de los niveles del factor *intra-sujetos* y del factor *inter-sujetos* en el botón Gráficos del cuadro de diálogo de MR (Figura 13.3) y se accede al cuadro de diálogo de la Figura 13.4. En este cuadro se traslada el factor *intra* al cuadro **Eje horizontal** y el factor *inter* al cuadro **Líneas distintas**, y en el Visor se obtiene el gráfico que se ve en la Figura 13.8.

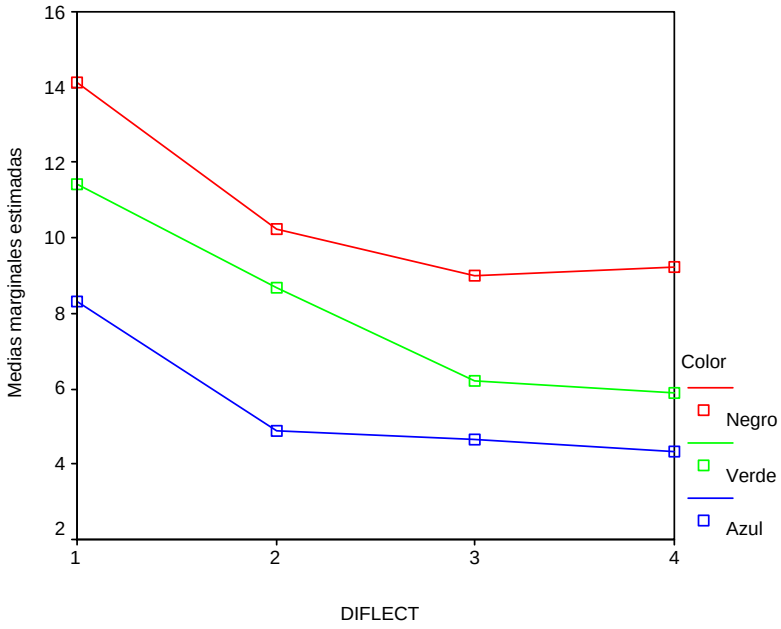


Figura 13.8 Gráfico de perfil de la interacción dificultad lectura-color de impresión

Se observa en el gráfico que efectivamente, no hay efecto de interacción, ya que la disminución del número de errores detectados baja gradualmente en los niveles del factor intra y es independiente del color en que esté impresa la página. Esta ausencia de interacción ya se ha comprobado analíticamente mediante la prueba de los efectos mostrada en la Tabla 13.6. También en el gráfico se puede ver que entre los niveles del factor inter hay diferencias en los promedios, y al menos una de estas diferencias resulta significativa según el resultado de la prueba de los efectos inter-sujetos de la Tabla 13.6. Para ver entre qué niveles de los factores de dan las diferencias se deben realizar las pertinentes comparaciones múltiples.

13.3.3 Comparaciones múltiples

La comparación de los niveles del factor *inter-sujetos* se puede realizar mediante las opciones del cuadro de diálogo **Post hoc**, que ya hemos visto en el capítulo 11. Para las comparaciones del factor intra-sujetos y de cada nivel del factor *inter* dentro de cada nivel del factor intra, se utiliza la opción **Comparar efectos principales** del cuadro de diálogo **Opciones** (ver Figura 13.6). Para ello se selecciona tanto el factor **diflect** como **diflect* color** en la lista **Factores e interacciones de los factores** y se traslada a la lista **Mostrar medias para**. Luego se marca la opción **Comparar efectos principales** y se selecciona la opción **Bonferroni** en el menú desplegable **Ajuste del intervalo de confianza**

(para el control de la tasa de error). Hechas estas elecciones se vuelve al cuadro de diálogo principal, y en lugar del botón **Aceptar** se pulsa **Pegar**. Luego, en la ventana de sintaxis, se modifica la línea

/ EMMEANS = TABLES (diflect* color)

añadiendo lo siguiente:

COMPARE (color) ADJ(BONFERRONI).

La primera tabla es la de la comparación entre los niveles del factor intra, tal como puede verse en la Tabla 13.8. En esta tabla se observa que hay diferencias significativas entre el nivel 1 y los niveles 2, 3 y 4, y entre el nivel 2 y los niveles 3 y 4. No hay diferencias entre los niveles 3 y 4.

Tabla 13.8 Comparaciones entre los niveles del factor *diflect*

Comparaciones por pares

Medida: MEASURE_1

(I) DIFLECT	(J) DIFLECT	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 % para diferencia [a]	
					Límite inferior	Límite superior
1	2	3,370	,390	,000	2,248	4,492
	3	4,667	,420	,000	3,458	5,875
	4	4,815	,414	,000	3,624	6,005
2	1	-3,370	,390	,000	-4,492	-2,248
	3	1,296	,422	,031	8,335E-02	2,509
	4	1,444	,376	,005	,362	2,526
3	1	-4,667	,420	,000	-5,875	-3,458
	2	-1,296	,422	,031	-2,509	-8,335E-02
	4	,148	,430	1,000	-1,087	1,383
4	1	-4,815	,414	,000	-6,005	-3,624
	2	-1,444	,376	,005	-2,526	-,362
	3	-,148	,430	1,000	-1,383	1,087

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

La otra tabla, obtenida por el añadido mencionado en la sintaxis del procedimiento, compara por pares los niveles del factor **color** para cada nivel del factor **diflect**. Se observa en el Tabla 13.9, que en todos los niveles del factor *intra*, hay diferencias significativas entre al menos dos niveles del factor *inter* (si el efecto de la interacción hubiera sido significativo, habría habido diferencias en los niveles del factor *inter* en alguno de los niveles del factor *intra*, pero no en todos). Así, por ejemplo, en el primer nivel de **diflect** hay diferencias entre todos los colores; en el segundo nivel no hay diferencias entre verde y negro, pero sí entre las restantes combinaciones. En el tercer nivel no diferencias entre azul y verde, pero sí entre todas las restantes combinaciones. Y, por último, en el cuarto nivel, no hay diferencias entre verde y azul, pero sí entre todas las demás comparaciones.

Tabla 13.9. Comparaciones entre los niveles del factor color en cada nivel del factor diflect

Comparaciones por pares

Medida: MEASURE_1

DIFLECT	(I) COLOR	(J) COLOR	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 % para diferencia [a]	
						Límite inferior	Límite superior
1	Negro	Verde	2,667	,929	,025	,277	5,056
		Azul	5,778	,929	,000	3,388	8,167
	Verde	Negro	-2,667	,929	,025	-5,056	-,277
		Azul	3,111	,929	,008	,721	5,501
	Azul	Negro	-5,778	,929	,000	-8,167	-3,388
		Verde	-3,111	,929	,008	-5,501	-,721
2	Negro	Verde	1,556	,841	,230	-,610	3,721
		Azul	5,333	,841	,000	3,168	7,499
	Verde	Negro	-1,556	,841	,230	-3,721	,610
		Azul	3,778	,841	,000	1,613	5,943
	Azul	Negro	-5,333	,841	,000	-7,499	-3,168
		Verde	-3,778	,841	,000	-5,943	-1,613
3	Negro	Verde	2,778	,858	,011	,569	4,987
		Azul	4,333	,858	,000	2,124	6,542
	Verde	Negro	-2,778	,858	,011	-4,987	-,569
		Azul	1,556	,858	,247	-,653	3,764
	Azul	Negro	-4,333	,858	,000	-6,542	-2,124
		Verde	-1,556	,858	,247	-3,764	,653
4	Negro	Verde	3,333	,895	,003	1,031	5,636
		Azul	4,889	,895	,000	2,586	7,191
	Verde	Negro	-3,333	,895	,003	-5,636	-1,031
		Azul	1,556	,895	,285	-,747	3,858
	Azul	Negro	-4,889	,895	,000	-7,191	-2,586
		Verde	-1,556	,895	,285	-3,858	,747

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

Las comparaciones por pares de los niveles del factor *inter* (**color**), se obtienen en el botón **Post hoc**. Para los datos que estamos manejando se han seleccionado la prueba de *Tukey* y de *Scheffé*, y el resultado se ve en la Tabla 13.10.

Tabla 13.10 Comparaciones múltiples del factor *color*

Medida: MEASURE_1

	(I) COLOR	(J) COLOR	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior
DHS de Tukey	Negro	Verde	2,58*	,63	,001	1,00	4,16
		Azul	5,08*	,63	,000	3,50	6,66
	Verde	Negro	-2,58*	,63	,001	-4,16	-1,00
		Azul	2,50*	,63	,002	,92	4,08
	Azul	Negro	-5,08*	,63	,000	-6,66	-3,50
		Verde	-2,50*	,63	,002	-4,08	-,92
Scheffe	Negro	Verde	2,58*	,63	,002	,93	4,23
		Azul	5,08*	,63	,000	3,43	6,73
	Verde	Negro	-2,58*	,63	,002	-4,23	-,93
		Azul	2,50*	,63	,002	,85	4,15
	Azul	Negro	-5,08*	,63	,000	-6,73	-3,43
		Verde	-2,50*	,63	,002	-4,15	-,85

Basado en las medias observadas.

*. La diferencia de medias es significativa al nivel ,05.

13.4 Modelo de dos factores, ambos con medidas repetidas

Supongamos ahora que además del nivel de dificultad de la página se añade otro factor *intra-sujetos* referente al contenido de dicha página, con dos niveles: contenido literario (L) y contenido científico (C), y en consecuencia todos los sujetos deben de leer ambos contenidos en todos los niveles de dificultad. Los datos, pues, podrían ser los siguientes:

	diflec1		diflec2		diflec3		diflec4	
Sujetos	L	C	L	C	L	C	L	C
1	14	13	12	12	7	8	6	7
2	15	10	10	10	9	5	9	7
3	16	9	8	8	11	5	9	8
4	13	10	11	8	8	7	9	8
5	16	12	12	9	7	5	12	5
6	16	13	10	8	8	10	11	5
7	14	10	13	7	12	6	10	4
8	12	12	8	9	11	6	7	4
9	11	14	8	7	8	4	10	5

Para introducir este tipo de diseños en el **Editor de datos** se hace como se muestra en la Figura 13.9.

1 : diflec1l | 14

	diflec1l	diflec1c	diflec2l	diflec2c	diflec3l	diflec3c	diflec4l	diflec4c
1	14	13	12	12	7	8	6	7
2	15	10	10	10	9	5	9	7
3	16	9	8	8	11	5	9	8
4	13	10	11	8	8	7	9	8
5	16	12	12	9	7	5	12	5
6	16	13	10	8	8	10	11	5
7	14	10	13	7	12	6	10	4
8	12	12	8	9	11	6	7	4
9	11	14	8	7	8	4	10	5

Figura 13.9 Datos en el Editor de un diseño de dos factores, ambos con medidas repetidas

Para trabajar con este tipo de datos, hay que, previamente, nombrar los dos factores en el cuadro de diálogo de la Figura 13.2, uno **diflec** (dificultad lectura) con 4 niveles y otro **conten** (contenido) con dos niveles. Una vez definidos, el cuadro de diálogo principal es el que se muestra en la figura 13.10.

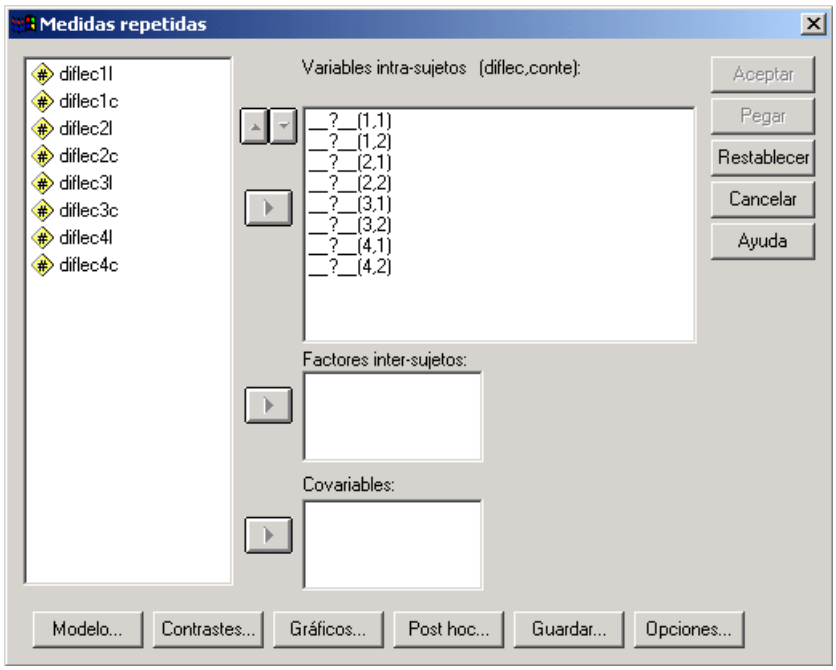


Figura 13.10 Cuadro de diálogo de MR con dos factores intra-sujetos

En la lista de **Variables intra-sujetos** están ya ubicados los niveles del primer factor definido y del segundo: el primer valor del primer paréntesis se corresponde con el primer nivel del primer factor, y el segundo valor se corresponde con el primer valor del segundo factor. El orden de los factores dependerá del orden en que se hayan definido en el cuadro de definición de factores. Como en nuestro caso se ha definido de acuerdo al orden en que aparecen las variables en el archivo, se marcan todas las variables y se trasladan a la lista **Variables intra-sujetos**. Sin opciones adicionales, ni gráficos, ni comparaciones múltiples, el análisis básico muestra en el Visor las tablas que se ven en la Tabla 13.11

Tabla 13.11 Tablas con los análisis básicos del diseño de dos factores con medidas repetidas en ambos

Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación
DIFLEC	Traza de Pillai	,963	52,649	3,000	6,000	,000
	Lambda de Wilks	,037	52,649	3,000	6,000	,000
	Traza de Hotelling	26,324	52,649	3,000	6,000	,000
	Raíz mayor de Roy	26,324	52,649	3,000	6,000	,000
CONTE	Traza de Pillai	,739	22,612	1,000	8,000	,001
	Lambda de Wilks	,261	22,612	1,000	8,000	,001
	Traza de Hotelling	2,826	22,612	1,000	8,000	,001
	Raíz mayor de Roy	2,826	22,612	1,000	8,000	,001
DIFLEC * CONTE	Traza de Pillai	,448	1,622	3,000	6,000	,281
	Lambda de Wilks	,552	1,622	3,000	6,000	,281
	Traza de Hotelling	,811	1,622	3,000	6,000	,281
	Raíz mayor de Roy	,811	1,622	3,000	6,000	,281

Prueba de esfericidad de Mauchly

Medida: MEASURE_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Significación	Epsilon		
					Greenhouse -Geisser	Huynh-Feldt	Límite-inf
DIFLEC	,788	1,601	5	,902	,892	1,000	
CONTE	1,000	,000	0	,	1,000	1,000	1
DIFLEC * CONTE	,586	3,594	5	,612	,807	1,000	

Pruebas de efectos intra-sujetos.

Medida: MEASURE_1

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
DIFLEC	Esfericidad asumida	324,042	3	108,014	40,295	,000
	Greenhouse-Geisser	324,042	2,675	121,123	40,295	,000
	Huynh-Feldt	324,042	3,000	108,014	40,295	,000
	Límite-inferior	324,042	1,000	324,042	40,295	,000
Error(DIFLEC)	Esfericidad asumida	64,333	24	2,681		
	Greenhouse-Geisser	64,333	21,402	3,006		
	Huynh-Feldt	64,333	24,000	2,681		
	Límite-inferior	64,333	8,000	8,042		
CONTE	Esfericidad asumida	120,125	1	120,125	22,612	,001
	Greenhouse-Geisser	120,125	1,000	120,125	22,612	,001
	Huynh-Feldt	120,125	1,000	120,125	22,612	,001
	Límite-inferior	120,125	1,000	120,125	22,612	,001
Error(CONTE)	Esfericidad asumida	42,500	8	5,313		
	Greenhouse-Geisser	42,500	8,000	5,313		
	Huynh-Feldt	42,500	8,000	5,313		
	Límite-inferior	42,500	8,000	5,313		
DIFLEC * CONTE	Esfericidad asumida	7,486	3	2,495	,774	,520
	Greenhouse-Geisser	7,486	2,421	3,092	,774	,498
	Huynh-Feldt	7,486	3,000	2,495	,774	,520
	Límite-inferior	7,486	1,000	7,486	,774	,405
Error(DIFLEC*CONTE)	Esfericidad asumida	77,389	24	3,225		
	Greenhouse-Geisser	77,389	19,371	3,995		
	Huynh-Feldt	77,389	24,000	3,225		
	Límite-inferior	77,389	8,000	9,674		

Pruebas de los efectos inter-sujetos

Medida: MEASURE_1

Variable transformada: Promedio

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Intercept	6290,681	1	6290,681	2461,571	,000
Error	20,444	8	2,556		

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	DIFLEC	CONTE	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significac
DIFLEC	Lineal		275,625	1	275,625	136,957	
	Cuadrático		48,347	1	48,347	15,931	
	Cúbico		6,944E-02	1	6,944E-02	,023	
Error(DIFLEC)	Lineal		16,100	8	2,012		
	Cuadrático		24,278	8	3,035		
	Cúbico		23,956	8	2,994		
CONTE		Lineal	120,125	1	120,125	22,612	
Error(CONTE)		Lineal	42,500	8	5,312		
DIFLEC * CONTE	Lineal	Lineal	2,336	1	2,336	,590	
	Cuadrático	Lineal	3,125	1	3,125	1,667	
	Cúbico	Lineal	2,025	1	2,025	,528	
Error(DIFLEC*CONTE)	Lineal	Lineal	31,689	8	3,961		
	Cuadrático	Lineal	15,000	8	1,875		
	Cúbico	Lineal	30,700	8	3,837		

La primera tabla, como en los casos anteriores, contiene los contrastes multivariados, según los cuales, hay diferencias significativas en los dos factores *intra-sujetos*, pero no en lo referente a la interacción. La siguiente tabla ofrece la prueba de esfericidad de *Mauchly*, según la cual no se puede rechazar dicha hipótesis, ni en el factor dificultad de lectura, ni en la interacción entre este factor y el de contenido. Respecto al factor contenido, no tiene sentido su cálculo puesto que sólo tiene dos niveles y por tanto una sola covarianza.

Puesto que la prueba de esfericidad señala la igualdad de la matriz de varianzas-covarianzas, deberemos fijarnos en los contrastes univariados (pruebas de los efectos *intra-sujetos*). Según esta prueba, los efectos de ambos factores son significativos, pero no lo es la interacción.

Por último, se muestra la prueba de contrastes de las tendencias de los dos factores. Obviamente, la única tendencia del factor **contenido** será la lineal, dado que sólo tiene dos niveles.

Al igual que en los diseños anteriores, es conveniente confeccionar un gráfico de perfil y también realizar las comparaciones múltiples para comparar los efectos principales. Si quisiéramos realizar las comparaciones múltiples de cada nivel de un factor para cada nivel del otro (efectos simples) se tienen que recurrir a la sintaxis y hacer la modificación ya comentada en los dos diseños anterior.

Sugerimos al lector que realice tanto los gráficos como las comparaciones y lo coteje con la gráfica de la Figura 13.11 y las tablas que se muestran en la Tabla 13.12.

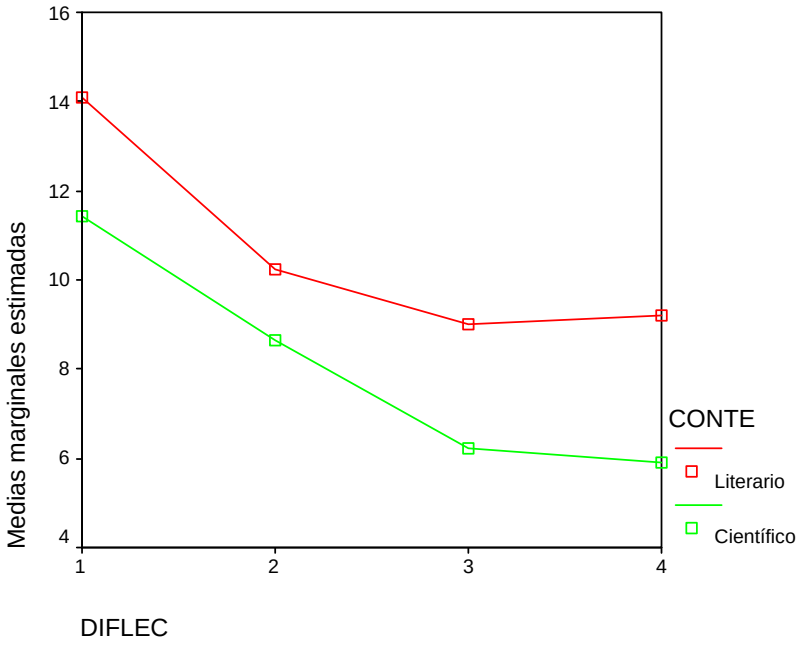


Figura 13.11 Gráfico de perfil de *dificultad lectura* por *contenido*

Tabla 13.12. Tablas con las comparaciones para cada efecto principal y los efectos simples. La primera tabla muestra las estimaciones de las medias para los cruces de los niveles

Estimaciones

Medida: MEASURE_1

DIFLEC	CONTE	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
1	1	14,111	,611	12,702	15,520
	2	11,444	,580	10,107	12,782
2	1	10,222	,641	8,745	11,700
	2	8,667	,527	7,451	9,882
3	1	9,000	,624	7,562	10,438
	2	6,222	,619	4,796	7,649
4	1	9,222	,619	7,796	10,649
	2	5,889	,539	4,647	7,131

Comparaciones por pares

Medida: MEASURE_1

(I) DIFLEC	(J) DIFLEC	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 % para diferencia [a]	
					Límite inferior	Límite superior
1	2	3,333*	,479	,001	1,668	4,999
	3	5,167*	,527	,000	3,333	7,000
	4	5,222*	,434	,000	3,713	6,732
2	1	-3,333*	,479	,001	-4,999	-1,668
	3	1,833	,607	,099	-,277	3,944
	4	1,889	,605	,085	-,217	3,995
3	1	-5,167*	,527	,000	-7,000	-3,333
	2	-1,833	,607	,099	-3,944	,277
	4	5,556E-02	,598	1,000	-2,024	2,135
4	1	-5,222*	,434	,000	-6,732	-3,713
	2	-1,889	,605	,085	-3,995	,217
	3	-5,556E-02	,598	1,000	-2,135	2,024

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

Comparaciones por pares

Medida: MEASURE_1

(I) CONTE	(J) CONTE	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 % para diferencia [a]	
					Límite inferior	Límite superior
1	2	2,583*	,543	,001	1,331	3,836
2	1	-2,583*	,543	,001	-3,836	-1,331

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

Comparaciones por pares

Medida: MEASURE_1

DIFLEC	(I) CONTE	(J) CONTE	Diferencia entre medias (I-J)	Error típ.	Significación [a]	Intervalo de confianza al 95 % para diferencia [a]	
						Límite inferior	Límite superior
1	1	2	2,667*	,986	,027	,393	4,940
	2	1	-2,667*	,986	,027	-4,940	-,393
2	1	2	1,556	,729	,065	-,125	3,236
	2	1	-1,556	,729	,065	-3,236	,125
3	1	2	2,778*	,983	,022	,511	5,044
	2	1	-2,778*	,983	,022	-5,044	-,511
4	1	2	3,333*	,928	,007	1,193	5,473
	2	1	-3,333*	,928	,007	-5,473	-1,193

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

14. Análisis de correlación y regresión

14.1 Introducción

En el capítulo de análisis de datos categóricos hemos visto cómo cuantificar la relación entre variables nominales y/o ordinales, pero nada se ha dicho hasta el momento de cómo cuantificar la relación que se puede dar entre variables cuantitativas o de escala, siguiendo la terminología de SPSS. En el ámbito de las Ciencias Sociales y de la Salud en muchas ocasiones se abordan estudios cuyas variables están medidos a este nivel y por ello es útil disponer de índices que permitan cuantificar la relación. En este capítulo nos vamos a circunscribir únicamente a las relaciones de tipo lineal, es decir a las relaciones que geométricamente podrían representarse mediante una línea recta en un eje cartesiano.

La manera más directa de observar si hay o no relación entre dos variables y de qué tipo es mediante el diagrama de dispersión. En los diferentes diagramas de las Figuras 14.1 se pueden ver algunas de las que se pueden dar entre dos variables.

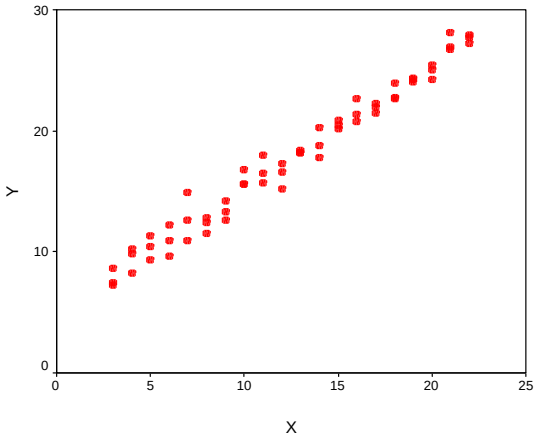


Figura 14.1(a)

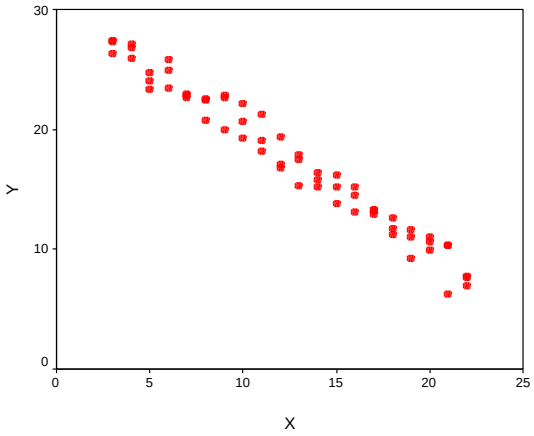


Figura 14.1(b)

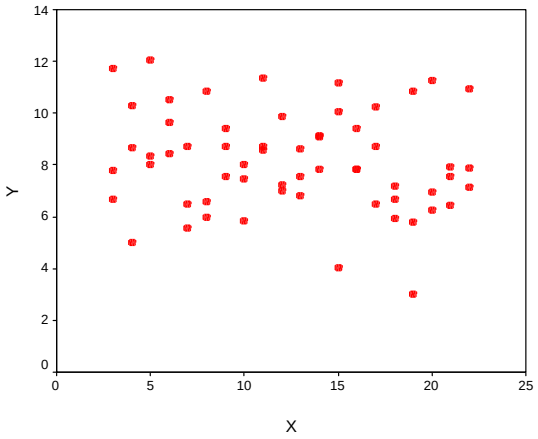


Figura 14.1(c)

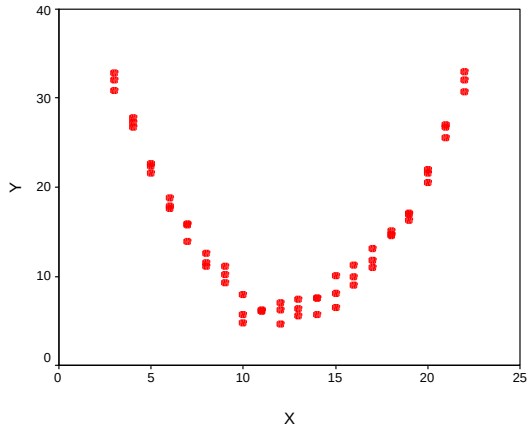


Figura 14.1(d)

La Figura 14.1 (a) representa una relación positiva (o directa) de tipo lineal, en la cual a valores bajos de la variable X corresponden valores bajos de la variable Y, y a valores altos de X corresponden valores altos de Y. La Figura 14.1 (b) representa una relación negativa (o inversa) de tipo lineal, en la cual a valores bajos de X corresponden valores altos de Y y viceversa. En los datos representados en la Figura 14.1 (c) la nube de puntos es muy dispersa en todos el rango de valores de X e Y, lo que indica ausencia de relación lineal. Y, por último, en la Figura 14.1 (d) hay relación entre las variables, pero su forma es parabólica y no lineal.

Una vez formada la idea gráfica de la relación entre las variables es preciso disponer de índices que permitan su cuantificación. Estos índices suelen denominarse coeficientes de correlación, y además de cuantificar la relación lineal entre dos variables permite determinar el grado de ajuste de una nube de puntos a una línea recta.

Los datos sobre los que vamos a trabajar son los que se muestran en el Cuadro 14.1 , y representan un estudio (ficticio) para determinar si la destreza manual (variable D) y la atención para la percepción de pequeños detalles (variable A) están relacionados con la productividad (variable P) de los operarios de una empresa de manufacturas. Para dicho estudio se dispone de una muestra de 90 operarios a quienes se administra una prueba de destreza manual y otra de atención a los detalles; por otro lado, se registra el número de piezas ensambladas por hora.

Cuadro 14.1. Datos de *Productividad* (P) de 90 operarios, y sus puntuaciones en una prueba de *Destreza manual* (D) y otra de *Atención a los detalles* (A)

P	D	A	P	D	A	P	D	A	P	D	A	P	D	A	P	D	A
10	54	39	11	44	70	14	59	73	9	48	48	7	48	39	10	49	48
9	41	30	9	39	49	13	52	58	8	38	48	9	41	42	10	38	45
9	43	51	10	44	49	11	60	54	9	42	59	10	56	52	7	40	45
10	55	48	7	48	52	11	55	61	8	52	35	10	50	39	7	47	44
13	57	60	13	50	62	11	67	48	3	42	13	12	58	63	9	63	48
11	50	52	6	46	52	6	35	35	6	21	35	8	46	52	11	66	48
8	49	53	13	64	54	13	64	59	10	36	63	11	50	64	9	59	44
9	43	57	12	46	77	9	41	56	8	42	43	11	60	48	12	51	45
11	52	62	10	57	57	7	54	45	12	55	50	9	53	39	8	51	40
9	61	46	11	70	42	7	52	33	8	41	51	9	37	44	6	43	40
9	39	51	12	40	54	8	36	44	12	69	58	8	49	42	7	34	47
10	53	56	9	57	58	13	64	58	11	47	52	9	33	48	14	60	52

9	42	46	11	34	47	10	55	58	10	55	49	9	44	46	8	62	39
10	49	42	8	49	33	9	57	39	8	43	45	6	24	42	10	54	54
10	51	46	10	41	53	9	40	44	13	52	51	7	46	32	11	51	51

14.2 Correlación lineal simple

Se dice que dos variables correlacionan o están relacionadas cuando ambas varían de forma conjunta. Esta variación conjunta se observa en el diagrama de dispersión, pero para determinar la magnitud se dispone de una serie de índices, denominados genéricamente coeficientes de correlación, que se emplean según sea el tipo de variables sobre las que se cuantifica la relación.

Para acceder en SPSS a estos índices, se sigue la secuencia

Analizar → Correlaciones → Bivariadas...

y se muestra el cuadro de diálogo de la Figura 14.2.

Figura 14.2 Cuadro de diálogo de Correlaciones bivariadas

Como siempre, en la lista de la izquierda se muestra las variables del archivo, o del conjunto de variables que hubiéramos definido. Después de seleccionar las variables y pasarlas a la lista Variables se señala el coeficiente de correlación que se quiere calcular. Los tres coeficientes de que se dispone son los siguientes:

- **Coeficiente de correlación de Pearson**. Es el adecuado para estudiar la relación lineal entre pares de variables cuantitativas o variables de escala (en terminología de SPSS). Se representa por r_{xy} y se puede expresar en términos de puntuaciones directas, diferenciales o típicas. En términos de estas últimas es un promedio de los productos de las puntuaciones típica de cada caso.

  En puntuaciones típicas: $$r_{xy} = \frac{\sum z_x z_y}{n}$$

En puntuaciones diferenciales: $$r_{xy} = \frac{\sum xy}{nS_xS_y}$$

En puntuaciones directas: $$r_{xy} = \frac{n\sum XY - \sum X\sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

Pearson toma valores entre –1 y +1; un valor de 1 indica una relación lineal perfecta positiva y un valor de –1 indica una relación lineal perfecta negativa. En ambos casos el diagrama de dispersión será una línea recta.

- **Tau-b de Kendall**. Este coeficiente de correlación se emplea para estudiar variables ordinales, y está basado en el número de inversiones y no inversiones entre los casos. La fórmula ya se ha presentado en el capítulo de *Análisis de datos categóricos*, en el epígrafe de *Datos ordinales*. También toma valores entre –1 y +1 y se interpreta de la misma forma que Pearson. Este índice se utiliza cuando el nivel de medida es ordinal y no se puede suponer que la distribución poblacional conjunta de las variables sea normal.
- **Spearman**. Conocido como la ***rho*** de Spearman, es el coeficiente de correlación de Pearson aplicado a variables ordinales, o a variables de escala que se han transformado en rangos. Toma valores entre –1 y +1 y se interpreta igual que Pearson. Este índice es una alternativa a Pearson cuando se incumple el supuesto de normalidad.

Junto a cada coeficiente, se realiza la **prueba de significación** para contrastar la hipótesis de que el coeficiente en la población es igual a cero. Para ello se emplea un estadístico tipificado, que en el caso de Pearson es:

$$T = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

que se distribuye según el modelo t de Student con *n-2* grados de libertad.

Además, se puede elegir si el contraste es bilateral, cuando no hay un criterio sobre la dirección de la relación, o unilateral, cuando sí hay un criterio sobre la dirección de la relación.

Por último, si se señala la opción **Marcar las correlaciones significativas** (opción activa por defecto) se obtiene el nivel crítico exacto asociado a cada coeficiente de correlación. Si se desactiva la opción, en el Visor se muestra un asterisco en aquellos coeficientes cuyo nivel crítico es menor de 0,05 y dos asteriscos si el nivel crítico es menor de 0,01.

Junto con los coeficientes de correlación, se puede obtener información adicional de estadísticos descriptivos (medias y desviaciones típicas), y las covarianzas y los productos cruzados en el botón **Opciones**.

Hechas las elecciones del coeficiente (en este caso sólo Pearson) y las Opciones de descriptivos, las tablas que se muestran en el Visor son los que se ven en la Tabla 14.1.

Tabla 14.1 Tablas de resultados de correlaciones bivariadas para el coeficiente de correlación de Pearson

Estadísticos descriptivos

	Media	Desviación típica	N
Productividad	9,54	2,00	90
Destreza manual	48,94	9,60	90
Atención para el detalle	48,83	9,79	90

Correlaciones

		Productividad	Destreza manual	Atención para el detalle
Productividad	Correlación de Pearson	1,000	,536**	,671**
	Sig. (bilateral)	,	,000	,000
	Suma de cuadrados y productos cruzados	356,493	917,132	1171,163
	Covarianza	4,006	10,305	13,159
	N	90	90	90
Destreza manual	Correlación de Pearson	,536**	1,000	,223*
	Sig. (bilateral)	,000	,	,035
	Suma de cuadrados y productos cruzados	917,132	8199,716	1866,642
	Covarianza	10,305	92,132	20,974
	N	90	90	90
Atención para el detalle	Correlación de Pearson	,671**	,223*	1,000
	Sig. (bilateral)	,000	,035	,
	Suma de cuadrados y productos cruzados	1171,163	1866,642	8538,475
	Covarianza	13,159	20,974	95,938
	N	90	90	90

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

*. La correlación es significante al nivel 0,05 (bilateral).

Cuando se correlaciona una variable consigo misma, SPSS muestra una coma en la celda correspondiente al nivel crítico. SPSS no puede calcular un coeficiente cuando en alguna de las variables todos los casos son perdidos (del sistema o se usuario).

Si hubiéramos marcado los coeficientes Tau-b de Kendal y Spearman, la tabla con los resultados son los que se muestran en la Tabla 14.2.

Tabla 14.2 Tabla con los resultados de los coeficientes Tau-b y Spearman

Correlaciones

			Productividad	Destreza manual	Atención para el detalle
Tau_b de Kendall	Productividad	Coeficiente de correlación	1,000	,359**	,454**
		Sig. (bilateral)	,	,000	,000
		N	90	90	90
	Destreza manual	Coeficiente de correlación	,359**	1,000	,152*
		Sig. (bilateral)	,000	,	,034
		N	90	90	90
	Atención para el detalle	Coeficiente de correlación	,454**	,152*	1,000
		Sig. (bilateral)	,000	,034	,
		N	90	90	90
Rho de Spearman	Productividad	Coeficiente de correlación	1,000	,526**	,641**
		Sig. (bilateral)	,	,000	,000
		N	90	90	90
	Destreza manual	Coeficiente de correlación	,526**	1,000	,229*
		Sig. (bilateral)	,000	,	,030
		N	90	90	90
	Atención para el detalle	Coeficiente de correlación	,641**	,229*	1,000
		Sig. (bilateral)	,000	,030	,
		N	90	90	90

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

*. La correlación es significativa al nivel 0,05 (bilateral).

14.3 Correlación parcial

Este procedimiento permite calcular la relación lineal entre dos variables excluyendo el efecto que una tercera variable pueda tener sobre ambas. El modo como se controla dicho efecto se explicará más adelante, en el procedimiento de análisis de regresión lineal.

Para acceder al procedimiento se sigue la secuencia

Analizar → Correlaciones → Parciales...

y se muestra el cuadro de diálogo de la Figura 14.3.

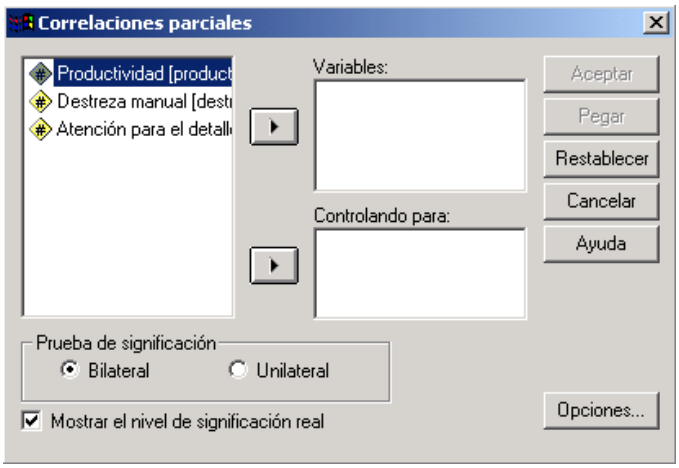


Figura 14.3 Cuadro de diálogo de Correlaciones parciales

A la lista **Variables** se trasladan las variables que queremos correlacionar, y a la lista **Controlando para** las variables cuyo efecto se desea controlar. En el botón **Opciones...** además de los estadísticos ya mencionados en las Opciones de correlaciones bivariadas, se pueden obtener las correlaciones de orden cero, es decir las correlaciones entre variables sin excluir el efecto de terceras variables.

El procedimiento Correlaciones parciales puede manejar una máximo de 400 variables de las que se puede ejercer el control sobre un máximo de 100. Las ecuaciones para obtener los coeficientes se pueden encontrar en numerosos textos de estadística (p.e. Cohen y Cohen, 1960; Hays, 1994), y su estructura es la siguiente:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$ Correlación parcial de orden uno

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{\left(1 - r_{14.3}^2\right)\left(1 - r_{24.3}^2\right)}}$$ Correlación parcial de orden dos

Para coeficientes de orden superior se sigue la misma lógica.

También se ofrece la prueba de significación en el Visor de resultados, mediante el siguiente estadístico:

$$T = \frac{r_{12.k}\sqrt{n - k - 2}}{\sqrt{1 - r_{12.k}^2}}$$

donde *n* es el número mínimo de casos con puntuaciones válidas en el conjunto de posibles correlaciones de orden cero y *k* es el número de variables controladas.

El tipo de contraste, **Bilateral** o **Unilateral** y la opción de **Mostrar el nivel de significación real**, es igual que lo ya explicado en correlaciones bivariadas.

Hecha la selección de las variables a correlacionar (*Productividad* y *Destreza*) y la de control (*Atención*), el resultado que se muestra en el visor es el de la Tabla 14.3.

Tabla 14.3 Resultados de los coeficientes de orden cero y parcial del procedimiento correlaciones parciales.

```
 P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S- -
Zero Order Partials
             PRODUCT   DESTREZA   ATENCION
PRODUCT       1,0000      ,5364      ,6713
             (    0)    (   88)    (   88)
             P= ,       P= ,000    P= ,000
DESTREZA        ,5364    1,0000      ,2231
             (   88)    (    0)    (   88)
             P= ,000    P= ,       P= ,035
ATENCION        ,6713      ,2231    1,0000
             (   88)    (   88)    (    0)
             P= ,000    P= ,035    P= ,
(Coefficient / (D.F.) / 2-tailed Significance)
" , " is printed if a coefficient cannot be computed
 P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S  -
Controlling for..    ATENCION
             PRODUCT   DESTREZA
PRODUCT       1,0000      ,5352
             (    0)    (   87)
             P= ,       P= ,000
DESTREZA        ,5352    1,0000
             (   87)    (    0)
             P= ,000    P= ,
(Coefficient / (D.F.) / 2-tailed Significance)
```

En este caso el influjo de la variable *Atención* sobre las variables *Productividad* y *Destreza* no afecta a la relación de orden cero entre ellas, ya que el valor del coeficiente de orden cero es 0,5364 y el valor de la correlación parcial es 0,5352, valor que también resulta significativamente distinto de cero.

14.4 Regresión lineal simple

Si la cuantía del coeficiente de correlación permite determinar que dos variables están relacionadas linealmente, es razonable pensar que si se conoce el comportamiento de una variables se pueda predecir el comportamiento de la otra, con la cual está relacionada aquélla. El instrumento que permite dicha predicción es el análisis de regresión, técnica estadística, indisociable de la correlación lineal, mediante la cual se puede explorar y cuantificar la relación entre dos o más variables, una de ellas denominada *variable dependiente* o *variable criterio* (Y) y

la/s otra/s denominada/s *variable/s independiente/s* o *variable/s predictora/s* (X_1, X_2,..., X_k).

14.4.1 La recta de regresión

Cuando se trata de la relación entre dos variables, ya se ha visto que el diagrama de dispersión es una buena aproximación para detectar si las variables se pueden ajustar más o menos bien a una línea recta. Ajustar un conjunto de puntos a una línea de modo que una vez construida dicha línea sirva para predecir el comportamiento de una variable. Se trata, pues de obtener una línea recta, que como cualquier recta se expresa como una función matemática simple del tipo:

$$\mathbf{Y_i = B_0 + B_1X_i}$$

donde el coeficiente B_0 (más conocido como intercepto) es el valor que toma la recta cuando X es igual a cero, y el coeficiente B_1, es la pendiente de la recta, o lo que es igual el cambio medio que se observa en Y por cada unidad que cambia X. Por ejemplo, la recta para ajustar los datos de Productividad en función de la Destreza manual es la que se puede ver en la Figura 14.4

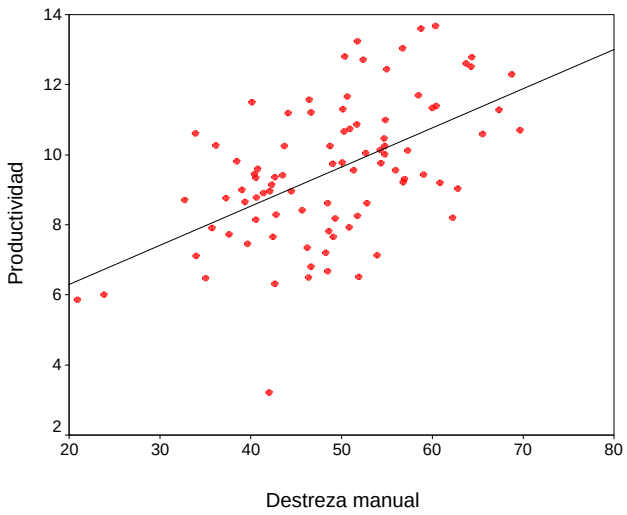


Figura 14.4 Diagrama de dispersión con la línea de regresión

$\mathbf{Y_i = 4{,}062 + 0{,}112X_i}$

Productividad = 4,062 + 0,112 · Destreza

La lectura de la recta nos informa de dos cuestiones importantes: en primer lugar, que la productividad aumenta un factor de 0,112 por cada aumento unitario que se da en la prueba de destreza manual; en segundo lugar, que la productividad de un operario que obtuviera un cero en la prueba de destreza sería, en promedio, de 4,062 piezas ensambladas a la hora. Este coeficiente (intercepto) es una información extrapolada del conjunto de datos, en el sentido de que no hay ningún operario en la muestra que obtenga un valor cero en destreza manual, razón por la cual es poco informativo y sobre todo engañoso, dado que hacemos predicciones fuera del rango de datos, y esto en el contexto del análisis de regresión es inadecuado por arriesgado (nada impide que fuera de ese rango de datos observados, la relación entre las variables pueda ser de un tipo deferente al lineal).

Una vez obtenida la recta, todas las predicciones de productividad que se hagan caerán sobre la propia recta, es decir, para un operario que en la prueba de destreza obtiene 40 puntos, se le pronosticará una productividad promedio de 8,542 piezas a la hora. Si observamos el Cuadro 14.1 con los datos vemos que en los tres casos que hay con una destreza de 40, una tiene una productividad de 7 otro de 9 y otro de 12 piezas a la hora; es decir, ninguno de los valores reales coincide con los valores observados, lo que supone que al efectuar pronósticos con la recta se cometen una serie de errores, que son las diferencias entre los valores pronosticados y los valores reales u observados.

14.4.2 Cálculo de los coeficientes de la recta

De todas la posibles rectas que se pueden construir para ajustar una nube de puntos, sólo una es la que mejor ajusta los datos al cumplir los siguientes criterios: el primero, que el promedio de los pronósticos coincida con el promedio de los valores de la variable dependiente o variable criterio, es decir, que el pronóstico sea insesgado; si llamamos Y_i' a los valores pronosticados por la recta de regresión, al ser insesgado, entonces $\overline{Y}' = \overline{Y}$; el segundo criterio es que la suma cuadrática de los errores (error = distancia vertical entre el valor observado y el valor pronosticado) sea mínima. Sólo una recta cumple simultáneamente con estos dos criterios, y los coeficientes de dicha recta se obtienen mediante la siguientes formulas (con puntuaciones directas primero y diferencias después):

$$B_1 = \frac{\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$B_1 = \frac{\sum xy}{\sum x^2}$$

$$B_0 = \overline{Y} - B_1\overline{X}$$

El coeficiente B_1 se puede obtener también a partir de otros estadísticos, como en la siguiente fórmula:

$$B_1 = r_{xy}\frac{S_y}{S_x}$$

14.4.3 Grado de ajuste de la recta a los datos

Siempre se puede encontrar una recta mínimo-cuadrática que ajuste una nube de puntos, cualquiera que sea la forma de la nube. Por lo tanto, más importante que encontrar esa recta es determinar si se ajuste bien o no a esos datos. Una medida del ajuste es el denominado **coeficiente de determinación**, R^2, que en el contexto del análisis de regresión lineal es el cuadrado del coeficiente de correlación de Pearson. Este coeficiente, que toma forzosamente valores entre 0 y 1 (0 cuando las variables no están relacionadas y 1 cuando lo están totalmente), especifica la ganancia que podemos obtener al predecir una variable usando la información de otra u otras variables con las que está relacionada. Dicho de otra manera, R^2, especifica qué parte de la variabilidad de la VD es atribuida a la variabilidad de la/s VI/s.

Para los datos de *Productividad*, *Destreza* y *Atención* que estamos manejando, sabemos que el coeficiente de correlación de Pearson entre Productividad y Destreza es 0,5364, y por tanto R^2 = 0,2877. Esto significa que al hacer pronósticos de la Productividad basándonos en la información de la Destreza, se mejora el pronóstico el 28,77%, respecto del pronóstico en el que sólo hubiéramos tomado en consideración la propia variable Productividad. Es decir, en términos de variabilidad asociada, el 28,77% de la varianza de la Productividad se debe a la varianza de la destreza. Como los pronósticos son una transformación lineal de la variable independiente X, entonces ese porcentaje de varianza de la variable dependiente Y es el porcentaje explicado por los pronósticos, mientras que el resto del porcentaje de la variabilidad (100 – 28,77 = 71,23), es el que se atribuye al error que se comete al pronosticar.

Se ve, pues, que la variabilidad de la VD se puede descomponer en dos variabilidades: la variabilidad de los pronósticos y la variabilidad de los errores o de los residuos, lo cual puede expresarse de la siguiente manera:

$$S_Y^2 = S_{PRONÓSTICOS}^2 + S_{RESIDUOS}^2$$

La proporción de la varianza de Y atribuible a la varianza de los pronósticos es el coeficiente de determinación:

$$R^2 = \frac{S_{PRONÓSTICOS}^2}{S_Y^2}$$

expresado en términos de suma de cuadrados, sería:

$$R^2 = \frac{\text{Suma de cuadrados de los pronósticos}}{\text{Suma de cuadrados total}}$$

o también:

$$R^2 = 1 - \frac{\text{Suma de cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

14.5 Análisis de regresión lineal simple

El procedimiento Regresión lineal de SPSS proporciona los coeficientes de la recta de regresión y su grado de ajuste de una manera sencilla. Para ilustrar el procedimiento vamos a obtener la ecuación de regresión de la Productividad (VD) respecto de la Destreza (VI). Para acceder al procedimiento se sigue la secuencia

Analizar → Regresión → Lineal...

y se muestra el cuadro de diálogo de la Figura14.5.

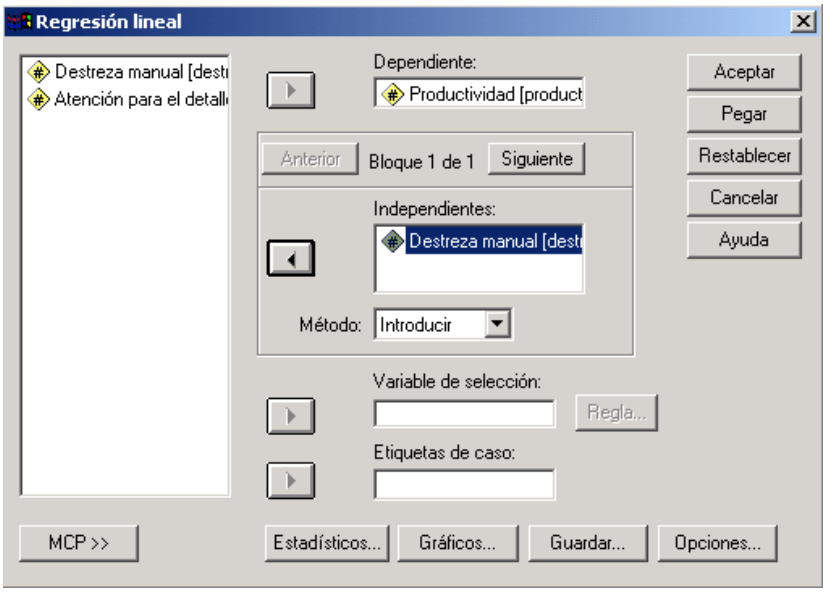


Figura 14.5 Cuadro de diálogo de Regresión lineal

Al cuadro **Dependiente** se traslado la VD y a la lista **Independientes** la variable independiente en el caso de la regresión lineal simple, o el grupo de variables independientes (en el caso de la regresión lineal múltiple, que se estudia posteriormente). Hecha la selección de las variables los resultados que se muestran en el Visor son los que se ven en las tablas de la Tabla 14.4.

Tabla 14.4 Resultados por defecto del procedimiento Regresión lineal con un solo predictor

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,536[a]	,288	,280	1,70

a. Variables predictoras: (Constante), Destreza manual

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	102,581	1	102,581	35,552	,000[a]
	Residual	253,912	88	2,885		
	Total	356,493	89			

a. Variables predictoras: (Constante), Destreza manual

b. Variable dependiente: Productividad

Coeficientes [a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	4,062	,935		4,343	,000
	Destreza manual	,112	,019	,536	5,963	,000

a. Variable dependiente: Productividad

En la primera tabla, denominada **Resumen del modelo**, se muestra el coeficiente de correlación de Pearson (0,536) y su cuadrado, el coeficiente de determinación (0,288). También se muestra el valor de R^2 corregido (0,280), corrección que se basa en el número de casos (n) y de variables predictoras (p), y cuyo valor se acerca a R^2 a medida que aumenta el número de casos de la muestra. Su expresión es:

$$R^2_{\text{CORREGIDA}} = R^2 - \frac{p\left(1 - R^2\right)}{\left(n - p - 1\right)}$$

Esta corrección se efectúa para evitar los valores altos de R^2 cuando hay pocas variables predictoras y pocos casos. Para nuestros datos, con una predictora y 90 casos, los valores son muy similares.

El error típico de estimación (1,70), es la raíz cuadrada de la media cuadrática residual (se puede decir que es la desviación típica de los residuos), y representa, como ya se ha dicho, la parte de la variabilidad de la VD que no es explicada por la recta de regresión. Su expresión es:

$$\text{Error típico de estimación} = S_e = \sqrt{\frac{\sum\left(Y_i - Y_i'\right)^2}{n - 2}}$$

Lógicamente, cuanto mejor sea el ajuste, menor será este error.

La siguiente tabla muestra el ANOVA aplicado a la regresión, e informa de si la relación que hay entre las variables es o no significativa. El estadístico F, contrasta la hipótesis nula de que el valor del coeficiente de correlación es cero en la población, que es lo mismo que decir que la pendiente de la recta de regresión es cero. Si el nivel crítico asociado a la F es menor de 0,05 se rechaza la hipótesis nula y se acepta que hay relación entre las variables. Como el valor de F en este caso es significativo ($F = 35{,}552$; $p < 0{,}05$) se concluye que sí existe relación entre la Destreza y la Productividad.

La última tabla es la de los coeficientes de regresión, y se muestran los coeficientes no estandarizados y los estandarizados, así como la prueba de significación de dichos coeficientes. Los *coeficientes no estandarizados* son la constante (B_0) y la pendiente (B_1), son los coeficientes que se obtienen cuando se ajusta una recta sobre un par de variables expresadas en puntuaciones directas, y su cálculo ya ha sido explicada. Por su parte, los *coeficientes estandarizados*, son los coeficientes que se obtienen cuando se ajusta la regresión sobre las puntuaciones típicas, y para obtenerlas se procede de la siguiente manera:

$$\beta_1 = B_1 \frac{S_x}{S_y}$$

El lector puede deducir, a partir de la fórmula ya expuesta, $B_1 = r_{xy} \frac{S_y}{S_x}$, que $\beta_1 = r_{xy}$. es decir, los coeficientes de regresión estandarizados son los coeficientes de correlación de orden cero entre cada variable predictora y la variable

dependiente. Estos coeficientes muestran mejor que los no estandarizados la importancia relativa de cada predictor dentro de la ecuación de regresión.

En la tabla de los coeficientes también se muestra las pruebas de significación de los mismos. Los estadísticos t_{B_0} y t_{B_1} para estas pruebas se obtienen dividiendo los coeficientes de regresión entre sus respectivos errores típicos, calculándose éstos del siguiente modo:

$$S_{B_0} = S_e \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}} \qquad \text{y} \qquad S_{B_1} = \frac{S_e}{\sum (X_i - \overline{X})^2}$$

siendo S_e el error típico de estimación o desviación típica de los errores de pronóstico. Los estadísticos se distribuyen según el modelo *t* de Student con *n-2* grados de libertad. En el caso de la regresión simple, el estadístico t es equivalente al estadístico F de la tabla del ANOVA (observe el lector que t^2 = F: $5{,}963^2$ = 35,552). Para los datos que estamos analizando, el nivel crítico (*Sig.*) del estadísticos t del coeficiente de regresión, permite concluir que es significativamente distinto de cero, es decir, que entre Destreza y Productividad hay una relación lineal significativa.

14.6 Análisis de regresión lineal múltiple

Mediante el análisis de regresión lineal múltiple se puede determinar cómo se comporta una variable a partir una combinación óptima de un grupo de variables predictoras. En este caso lo que se construye como mejor predicción no es una recta sino un hiperplano.

Para ilustrar el análisis vamos a incorporar un nuevo predictor, la "*Atención para percibir los detalles*" (Variable A en el cuadro 14.1). En estas condiciones, el diagrama de dispersión[6] tridimensional (dos ejes para los predictores y otro para el criterio), con el plano de regresión incorporado sería el que se muestra en la Figura 14.6. La ecuación de este plano de regresión, expresada para puntuaciones directas, junto con el estadístico de bondad de ajuste, que se muestran en la parte superior del gráfico, son los siguientes:

Productividad = -0,41 + 0,12·Atención para el detalle + 0,08·Destreza

R^2 = 0,61

La estructura de la ecuación es la misma que la de la regresión lineal simple, sólo que con más predictores, cuyos coeficientes de correlación se calculan de modo que hagan mínimas las diferencias cuadráticas entre los valores observados y los pronosticados por el modelo. En general, la ecuación de regresión tendría la forma:

$$Y' = B_0 + B_1X_1 + B_2X_2 + \ldots + B_nX_n$$ expresado en puntuaciones directas

[6] El diagrama de dispersión se ha realizado con la opción de gráficos interactivos de SPSS, que dispone de más posibilidades que la opción diagrama de dispersión de los gráficos normales.

$$z_{y'} = \beta_0 + \beta_1 z_{x_1} + \beta_2 z_{x_2} + \ldots + \beta_n z_{x_n}$$ expresado en puntuaciones típicas

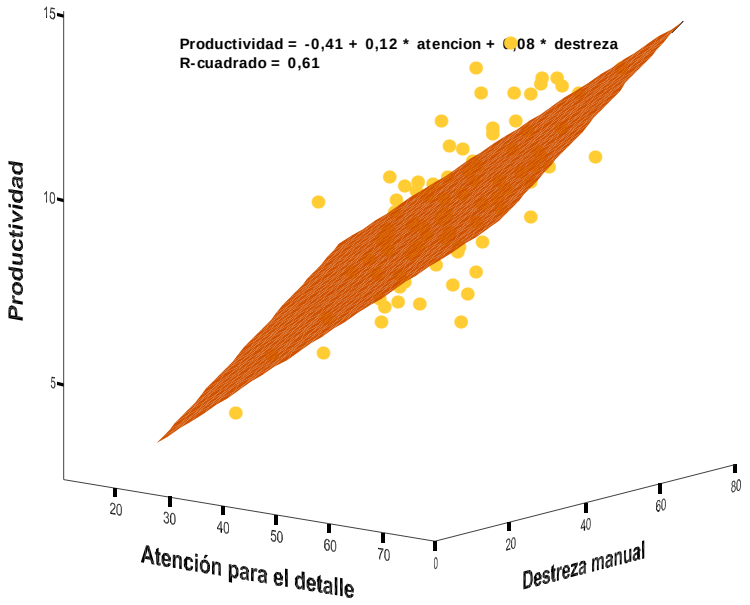


Figura 14.6 Diagrama de dispersión de Productividad sobre Atención para el detalle y Destreza manual

Los coeficientes B_i son los coeficientes no estandarizados y cuantifican el cambio que se produce en la variable dependiente por cada cambio unitario en ese predictor cuando el resto de los predictores permanecen fijos. Es decir, para nuestro ejemplo, cuando aumenta 1 punto en la prueba de "Atención", la productividad aumenta, en promedio, 0,12 puntos. Los coeficientes β_i son los coeficientes estandarizados, y cuantifican el cambio en términos de desviaciones típicas de la variable dependiente, por cada desviación típica que cambio ese predictor cuando el resto permanece constante. En este caso la ecuación en típicas sería:

$$Z'_{\text{Productividad}} = 0{,}580 \cdot Z_{\text{atención para el detalle}} + 0{,}407 \cdot Z_{\text{destreza manual}}$$

14.6.1 Grado de ajuste en la regresión lineal múltiple

Una vez obtenida la ecuación de regresión múltiple, es preciso saber si ajusta bien los datos, o lo que es igual si tiene una "buena" capacidad predictiva. Para responder a esta cuestión se dispone del denominado *coeficiente de correlación múltiple*, usualmente simbolizado por R, o también del cuadrado de este coeficiente, R^2, conocido como *coeficiente de determinación múltiple* (que expresa la proporción de varianza explicada).

El coeficiente de correlación múltiple cuantifica la correlación entre los valores pronosticados y los valores observados, aunque no es preciso calcular los valores

pronosticados, mediante la ecuación de regresión múltiple, y correlacionarlos con los observados. El valor de R puede obtenerse directamente, pues es la raíz cuadrada de una combinación lineal de los coeficientes de correlación de orden cero entre la variable dependiente y cada una de las variables predictoras, ponderados por los coeficientes de regresión parcial estandarizados (β_i). Su expresión es la siguiente:

$$R_{y.12...k} = \sqrt{\beta_1 r_{y1} + \beta_2 r_{y2} + ...\beta_k r_{yk}}$$

Puede verse esta igualdad con los datos que sirven de ejemplo.

$$0{,}780 = \sqrt{0{,}407 \times 0{,}5364 + 0{,}580 \times 0{,}6713}$$

donde 0,407 y 0,580 son los coeficientes de regresión estandarizados, como se puede ver en la tabla **Resumen del modelo** en la Tabla 14.5.

Obviamente, el coeficiente de correlación múltiple también puede expresarse a partir de los coeficientes de regresión no estandarizados, B_i, teniendo en cuenta que $\beta_i = B_i \frac{S_{X_i}}{S_Y}$.

14.6.2 Regresión lineal múltiple con SPSS

Para realizar un análisis de regresión lineal múltiple, en el cuadro de diálogo de Regresión lineal (ver Figura 14.5) se traslada las variable predictoras a la lista Variables independientes y la variable dependiente al lugar que le corresponde. Con el método de análisis que está definido por defecto (Introducir) las tablas que se muestran en el *Visor* de resultados son las que se ven en la Tabla 14.5.

Tabla 14.5. Resultados por defecto del procedimiento *Regresión lineal* con más de un predictor

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,780[a]	,608	,599	1,267

a. Variables predictoras: (Constante), Destreza manual, Atención para el detalle

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	216,733	2	108,366	67,457	,000[a]
	Residual	139,760	87	1,606		
	Total	356,493	89			

a. Variables predictoras: (Constante), Destreza manual, Atención para el detalle

b. Variable dependiente: Productividad

Coeficientes [a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-,409	,877		-,466	,642
	Atención para el detalle	,119	,014	,580	8,430	,000
	Destreza manual	8,485E-02	,014	,407	5,909	,000

a. Variable dependiente: Productividad

Según estos resultados, se concluye que el modelo de regresión se ajusta bien a los datos, dado que es capaz de reducir el error de predicción de la variable dependiente en casi el 61% (valor de R^2) cuando se toma en cuenta la información de los dos predictores. Además, los coeficientes de regresión son significativamente distintos de cero (ver los valores de la prueba t), mientras que el valor de la constante no es significativamente distinto de cero. A la misma conclusión se llega, en lo referente a la regresión, con la tabla del ANOVA de la regresión, en donde el estadístico F debido a la regresión indica que, efectivamente, ésta es significativa.

14.6.3 Información sobre estadísticos del procedimiento de regresión lineal

Además de la ecuación de regresión y el grado de ajuste del modelo, interesa disponer de la información sobre determinados estadísticos, tanto de cada variable por separado como de las relaciones entre ellas. Pulsando el botón Estadísticos del cuadro de diálogo de la Regresión Lineal (ver Figura 14.5), se accede al cuadro de diálogo que se muestra en la Figura 14.7.

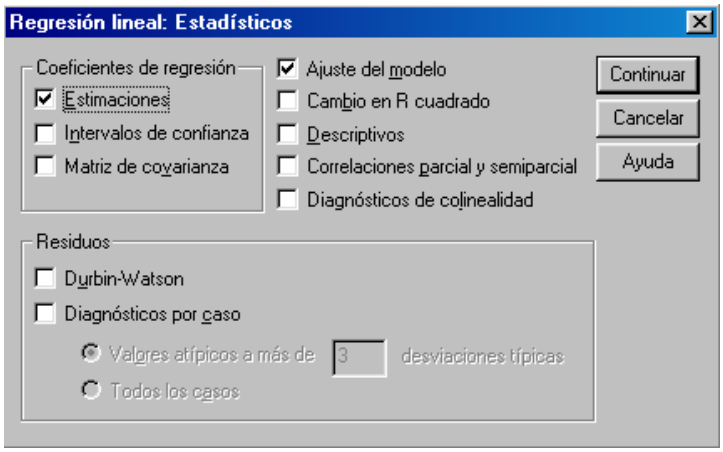


Figura 14.7 Cuadro de diálogo de *Estadísticos* de la *Regresión lineal*

Algunas de las opciones, además de las marcadas por defecto (**Estimaciones** y **Ajuste del modelo**), ya explicadas, son las siguientes:

- **Descriptivos**. Ofrece una tabla con la media, la desviación típica y el número de casos.
- **Matriz de covarianza**. Matriz de correlaciones y covarianzas entre los coeficientes de regresión del modelo.

- **Intervalos de confianza**. Se muestra en la tabla de los coeficientes y muestra los intervalos de confianza de cada uno de los coeficientes de regresión. Por defecto se calculan a una nivel de confianza de 0,95. Intervalos muy amplios denotan estimaciones poco precisas y son un aviso de problemas de colinealidad, concepto que se estudia posteriormente.
- **Correlaciones parcial y semiparcial**. Esta opción permite cuantificar la relación entre las variables predictoras y la variable dependiente, de dos maneras. La primera, la correlación parcial (en inglés, *partial correlation*) –ya explicada en este mismo capítulo-, expresa la relación entre una v. predictora y la v. dependiente después de eliminar sobre ambas el influjo del resto de vv. predictoras incluidas en la ecuación. Esta correlación se simboliza como $r_{y1.2}$, es decir, correlación parcial entre la variable independiente Y y la variable predictora X_1, eliminando sobre ambas el influjo de la variable predictora X_2.

  La correlación semiparcial, es la correlación entre dos variables, quitando sobre una de ellas el influjo de una tercera. Así, por ejemplo, la correlación semiparcial (en inglés *part correlation*) entre la variable dependiente Y y la predictora X_1, eliminando sobre esta última la influencia de la predictora X_2, se simboliza como $r_{y(1.2)}$.

Una relación que resulta de interés es la que permite expresar la proporción de varianza explicada en función tanto de las correlaciones parciales como de las semiparciales. Suponiendo solo dos predictores X_1 y X_2, la proporción de varianza no explicada se puede expresar como:

$$1 - R^2_{y.12} = \left(1 - r^2_{y1}\right)\left(1 - r^2_{y2.1}\right)$$

Es decir, la proporción de Y no explicada por las dos predictoras X_1 y X_2 es la proporción no explicada por X_1 veces la proporción no explicada por X_2 cuando X_1 permanece constante. De aquí se sigue que la proporción de varianza explicada es:

$$R^2_{y.12} = 1 - \left(1 - r^2_{y1}\right)\left(1 - r^2_{y2.1}\right)$$

En términos de correlación semiparcial, $R^2_{y.123}$ (una variable dependiente y tres predictoras) se escribiría del siguiente modo:

$$R^2_{y.12} = r^2_{y1} + r^2_{y(2.1)} + r^2_{y(3.12)}$$

Es decir, la proporción de varianza explicada es la suma de la proporción de varianza de Y explicada por X_1, más la proporción de varianza de Y explicada por X_2 cuando se elimina sobre ésta el influjo de las otras predictoras, más la proporción de varianza de Y explicada por X_3 cuando se elimina sobre ésta el influjo de las otras predictoras, y así sucesivamente.

Esta forma de enfocar el coeficiente de determinación múltiple R^2 nos permite determinar cuánta proporción de varianza de Y va siendo explicada por cada nuevo predictor que se añade a la ecuación de regresión.

En la Tabla 14.6 se presentan las tablas que contienen esta información.

Tabla 14.6. Tablas con Estadísticos descriptivos, Covarianza entre los coeficientes de regresión, y coeficientes de la ecuación de regresión, con la información de los coeficientes de correlación parcial y semiparcial

Estadísticos descriptivos

	Media	Desviación típ.	N
Productividad	9,54	2,001	90
Destreza manual	48,94	9,599	90
Atención para el detalle	48,83	9,795	90

Correlaciones de los coeficientes [a]

Modelo			Atención para el detalle	Destreza manual
1	Correlaciones	Atención para el detalle	1,000	-,223
		Destreza manual	-,223	1,000
	Covarianzas	Atención para el detalle	1,980E-04	-4,507E-05
		Destreza manual	-4,507E-05	2,062E-04

a. Variable dependiente: Productividad

Coeficientes [a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%		Correlaciones			Estadísticos de colinealidad	
		B	Error típ.	Beta			Límite inferior	Límite superior	Orden cero	Parcial	Semiparcial	Tolerancia	FIV
1	(Constante)	-,409	,877		-,466	,642	-2,151	1,334					
	Destreza manual	8,485E-02	,014	,407	5,909	,000	,056	,113	,536	,535	,397	,950	1,052
	Atención para el detalle	,119	,014	,580	8,430	,000	,091	,147	,671	,671	,566	,950	1,052

a. Variable dependiente: Productividad

El lector puede ver la relación mencionada entre el coeficiente de determinación múltiple R^2 y los coeficientes de correlación parcial y semiparcial.

$$R^2_{y.12} = 0,608 = 1-(1-0,671^2)(1-0,535^2) = 0,671^2 + 0,397^2$$

El resto de las opciones del cuadro de diálogo de estadístico tienen que ver con los supuestos del modelo de regresión lineal (*estadísticos de colinealidad* y *análisis de residuos*) y con el método de regresión por pasos una óptima ecuación de regresión (*cambio en R^2*)

14.6.4 Supuestos del modelo de regresión lineal

El primer supuesto que debe cumplir el modelo de regresión lineal es, precisamente, el de linealidad, es decir, que la variable dependiente puede expresarse como la suma de una constante, una combinación lineal de variables predictoras ponderadas por los coeficientes de regresión y un término error. Este supuesto previo puede comprobarse inspeccionando los diagramas de dispersión parciales entre la variable dependiente y cada una de las variables predictoras. Si

se incumple significa que los datos no pueden ser satisfactoriamente explicados por un modelo lineal y habría por tanto que ensayar el ajuste de otro tipo de modelos.

Si el modelo lineal es el adecuado, los cuatro supuestos básicos que tienen que cumplirse son los siguientes:

- **Independencia de los errores o residuos**. Los residuos (diferencias entre los valores observados y los pronosticados por el modelo) son una variable que se distribuye de manera aleatoria y deben ser independientes entre sí, es decir, no deben presentar autocorrelaciones.
- **Homocedasticidad**. La varianza de los residuos debe ser constante para cada valor de la variable independiente.
- **Normalidad.** Las distribuciones condicionadas de los residuos son normales con media cero.
- **No colinealidad**. Las variables predictoras no correlacionan entre sí. Si se incumple este supuesto se produce lo que se conoce como multicolinealidad.

14.6.4.1 Análisis de los residuos

En el cuadro de diálogo de Estadísticos hay dos opciones, en el apartado Residuos, que contrastar, por un lado, la Independencia y, por otro, permite obtener un listado con todos los residuos o bien sólo de aquellos que se desvían de cero más de un determinado número de desviaciones típicas (por defecto, está especificado a 3).

- **Independencia**. Para contrastar este supuesto SPSS calcula el estadístico de Durbin-Watson (1951) cuya expresión es:

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

donde ei son los residuos no estandarizados (Yi – Yi'). Este estadístico oscila entre 0 y 4 y toma el valor 2 cuando los residuos son independientes. Valores menores que 2 indican autocorrelación positiva y los mayores que 2 autocorrelación negativa. No obstante, se asume independencia con valores entre 1,5 y 2,5. Este estadístico, cuando se marca la opción, se muestra en la tabla correspondiente al Resumen del modelo, tal como puede verse en la tabla correspondiente en la Tabla 14.7. Para nuestro datos, el valor de este estadístico, 2,401, indica que no se puede rechazar la hipótesis de independencia.

Respecto del diagnóstico por caso, cuando hay residuos mayores (o menores) que el número de desviaciones típicas especificado, en el Visor se muestra un listado con dichos casos. Junto con esta tabla, se muestra otra con los estadísticos no sólo sobre los residuos sino también sobre los pronósticos, tanto estandarizados como no estandarizados. Para nuestros datos, no se ha generado ninguna tabla con los residuos que excedan las

tres desviaciones típicas porque no hay ninguno, tal como puede verse en la Figura 14.8, donde se ha representado los residuos tipificados frente a los pronósticos no estandarizados.

Tabla 14.7. Estadístico de Durbin-Watson y Estadísticos sobre los residuos

Resumen del modelo [b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,780[a]	,608	,599	1,267	2,401

a. Variables predictoras: (Constante), Atención para el detalle, Destreza manual

b. Variable dependiente: Productividad

Estadísticos sobre los residuos [a]

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	4,70	13,28	9,54	1,561	90
Residuo bruto	-3,20	3,19	,00	1,253	90
Valor pronosticado tip.	-3,099	2,400	,000	1,000	90
Residuo tip.	-2,524	2,517	,000	,989	90

a. Variable dependiente: Productividad

Para la representación gráfica de los residuos tipificados, primero hay que generarlos marcando la opción Correspondiente dentro del cuadro de diálogo Guardar. Posteriormente, se explican las opciones de este cuadro de diálogo con algo más de detalle.

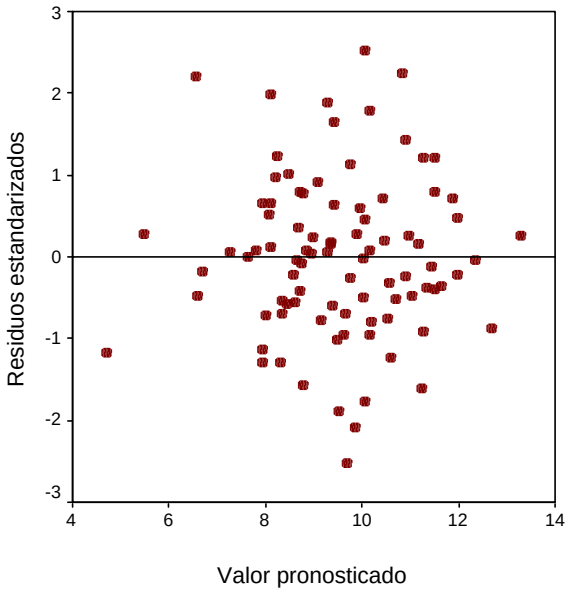


Figura 14.8 Diagrama de los residuos estandarizados frente a los valores pronosticados por la ecuación de regresión.

- **Homocedasticidad**. La mejor forma de determinar la homocedasticidad de los residuos es acudir a los gráficos que incorpora el procedimiento Regresión lineal. Para accede se pulsa el botón Gráficos y se muestra el cuadro de diálogo de la Figura 14.9

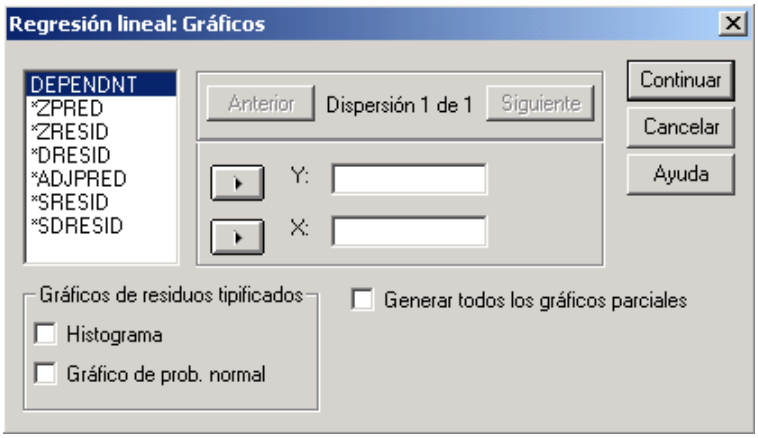


Figura 14.9 Cuadro de diálogo de *Gráficos* de *Regresión lineal*

De la lista de variables que se muestra en la izquierda, todas las precedidas por un asteriscos son variables generadas por SPSS, y se pueden crear con las opciones correspondientes del cuadro Guardar. Las variables son las siguientes:

- **DEPENDENT**. Variable dependiente de la ecuación de regresión.
- **ZPRED**. Pronósticos tipificados (con media 0 y desviación típica 1).
- **ZRESID**. Residuos tipificados. Si los residuos se distribuyen normalmente, es esperable que aproximadamente el 95% de los casos se encuentren en el rango –1,96 y +1,96. De esta manera es fácil identificar casos con residuos muy grandes.
- **DRESID**. Residuos eliminados o corregidos, que se obtienen al efectuar los pronósticos eliminando de la ecuación de regresión el caso sobre el que se realiza el pronóstico. Suelen ser útiles para detectar casos con influencia sobre la ecuación de regresión.
- **ADJPRED**. Pronósticos corregidos, que se obtienen eliminando el caso de la ecuación de regresión. Diferencias entre pronósticos y los pronósticos corregidos delatan casos de influencia.
- **SRESID**. Residuos estudentizados. Se obtienen tipificando los residuos, pero la desviación típica del cociente, se basa en la proximidad de cada caso respecto de su(s) media(s) en la(s) variable(s) independiente(s). Estos residuos están escalados en unidades de desviación típica, y se distribuyen según el modelo de probabilidad t de Student con n-p-1 grados de libertad (p se refiere al número de predictores). Con muestras grandes, aproximadamente el 95% de los residuos se encontrarían en el rango –2 y +2.
- **SDRESID**. Residuos corregidos estudentizados. Útiles para detectar casos de influencia.

De entre estas variables, las que permiten visualizar el supuesto de homocedasticidad son **ZRESID** y **ZPRED**, es decir, el diagrama de dispersión de los residuos tipificados frente a los pronósticos tipificados. Para realizar el gráfico, se pasa **ZRESID** al eje Y y **ZPRED** al eje X. El

aspecto de este gráfico, será similar al de la Figura 14.8, tal como puede verse en la Figura 14.9.

En este caso, el diagrama no presenta signos de heterocedasticidad (varianzas heterogéneas). En el caso de que sí lo hubiera, podría realizarse alguna transformación de la variable dependiente, pero teniendo en cuenta que puede variar la interpretación con el cambio de escala.

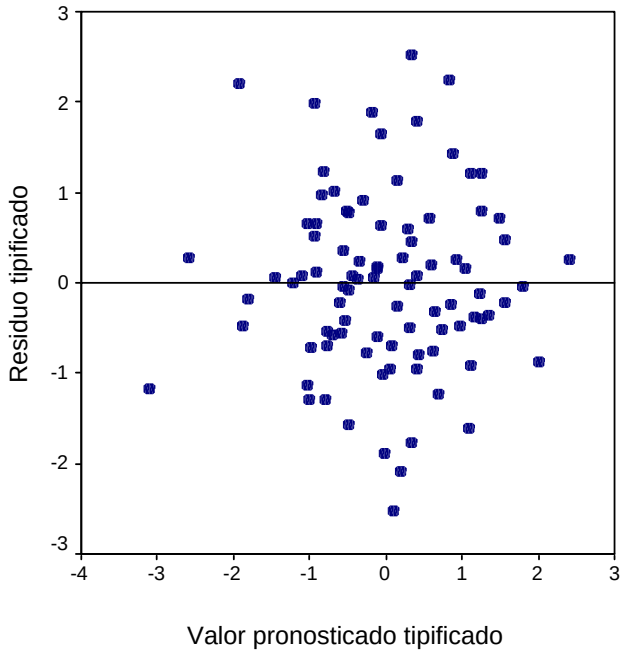
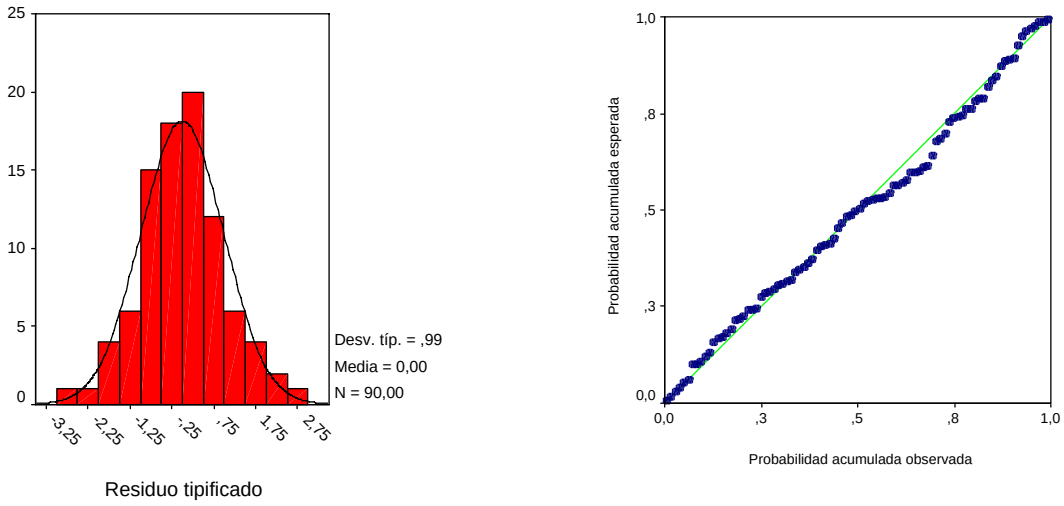


Figura 14.9 Diagrama de dispersión de los residuos tipificados frente a los pronósticos tipificados

- **Normalidad**. En el mismo cuadro de Gráficos de residuos tipificados, se puede obtener dos gráficos que permiten evaluar la normalidad de los residuos. En las Figuras 14.10 (a) y (b), se muestran estos dos gráficos.



Figuras 14.10. (a) Histograma de los residuos tipificados; (b) gráfico de probabilidad acumulada observada de los residuos frente a probabilidad acumulada esperada

El primero es un **Histograma de los residuos**, con la curva normal superpuesta. Se observa que la distribución es aproximadamente normal. El segundo es un gráfico de probabilidad normal, que compara la probabilidad cumulada observada, frente a la probabilidad acumulada esperada en caso de que la distribución fuera normal. También se ve en este gráfico que la distribución acumulada observada coincide con la distribución acumulada esperada. Estos dos procedimientos de evaluar la normalidad son gráficos, pero recordamos al lector que el procedimiento Explorar dispone de estadísticos que permite contrastar la normalidad de una distribución.

- **Linealidad**. Como se ha señalado al comienzo de este apartado de supuestos, la linealidad, o su ausencia, se detecta bien a través de los diagramas de dispersión. SPSS permite generar gráficos parciales de cada uno de los predictores con la variable dependiente, eliminando el efecto del resto de predictores. Estos diagramas no están basados en las puntuaciones originales de las dos variables escogidas, sino en los residuos de la regresión que se efectúa con el resto de predictores. Así, por ejemplo, el diagrama de regresión parcial entre "Productividad" y "Destreza", están representados los residuos de la regresión de "Productividad" sobre "Atención al detalle" y los residuos que resultan de regresar "Destreza" sobre "Atención al detalle". Por tanto, lo que se muestra, es la relación neta entre "Productividad" y "Destreza" eliminando el influjo que sobre ambas tiene la "Atención al detalle". Obviamente, si calculamos la correlación entre ambos residuos, se obtiene la correlación parcial entre "Productividad" y "Destreza".

  Cuando señalamos la opción **Generar todos los gráficos parciales** del cuadro de diálogo de Gráficos (ver Figura 14.9) se generan tantos gráficos parciales como variables independientes hay en la ecuación. En la Figura 14.11 se muestra el "Productividad" (v.d.) y "Atención al detalle".

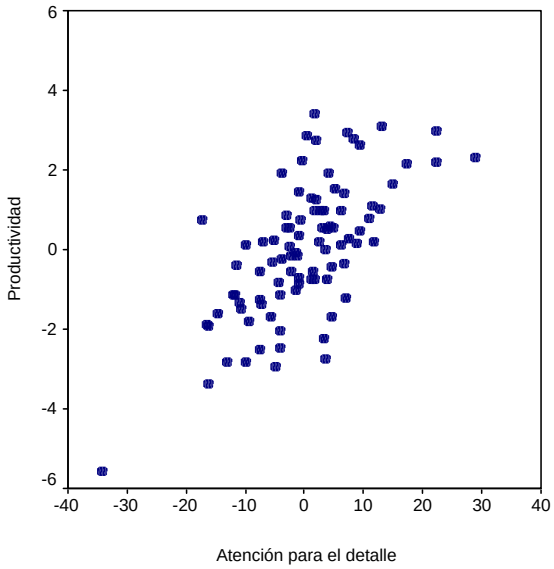


Figura 14.11. Gráfico de regresión parcial (Productividad por Atención al detalle)

Se ve que la relación entre ambas variables, una vez eliminada de ambas el influjo de "**Destreza**" es lineal y positiva.

- **Colinealidad**. Se dice que hay colinealidad cuando hay una correlación alta entre algunas, o todas, variables predictoras. En el caso de que una predictora sea una combinación lineal de otra, se habla de colinealidad perfecta, y en ese caso no es posible obtener los coeficientes de regresión parcial de la ecuación de regresión. Sin llegar a ese extremo, en numerosas ocasiones se produce el efecto de colinealidad entre predictores lo que produce un aumento de los residuos tipificados, que provoca ecuaciones de regresión con coeficientes muy inestables, en el sentido de que al añadir o quitar un caso, por ejemplo, se producen variaciones muy grandes de los coeficientes de regresión. Sin embargo, y eso el lector lo puede deducir de lo ya explicado, el coeficiente de determinación múltiple, R^2, no cambia en presencia de colinealidad. Para determinar si hay colinealidad, no existe un criterio estadístico formal, pero si se pueden dar algunas pautas:

  El estadístico F que evalúa el ajuste de la ecuación es significativo, pero no lo son ninguno de los coeficientes de regresión parcial.

  Los coeficientes de regresión estandarizados (Beta) adoptan en la misma ecuación valores por encima de 1 y por debajo de –1.

  Valores de tolerancia muy pequeños. La tolerancia de un predictor indica la proporción de varianza de esa variable no asociada al resto de los predictores incluidos en la ecuación.

  Los coeficientes de correlación estimados son muy grandes (por encima de 0,9).

  Además de estos indicios, SPSS dispone de algunos estadísticos que pueden ayudar a diagnosticar la presencia de colinealidad. Esta opción, Diagnósticos de colinealidad, se encuentra en el cuadro de diálogo de Estadísticos (ver

Figura 14.7). Al marcar esta opción, se muestra en el Visor una tabla con el Diagnóstico de colinealidad y, además, en la tabla con los coeficientes de regresión se incluye las tolerancias y sus inversos (FIV). En la Tabla 14.8, se pueden ver estas dos tablas.

Tabla 14.8. Tabla con la tolerancia y el FIV de los predictores y tabla con el diagnóstico de colinealidad.

Coeficientes [a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	-,409	,877		-,466	,642		
	Destreza manual	,085	,014	,407	5,909	,000	,950	1,052
	Atención para el detalle	,119	,014	,580	8,430	,000	,950	1,052

a. Variable dependiente: Productividad

Diagnósticos de colinealidad [a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza		
				(Constante)	Destreza manual	Atención para el detalle
1	1	2,955	1,000	,00	,00	,00
	2	,029	10,075	,00	,57	,65
	3	,015	13,816	1,00	,42	,35

a. Variable dependiente: Productividad

La *tolerancia* de un predictor se obtiene restando a 1 el coeficiente de determinación, R^2, que resulta de regresar esa variable sobre el resto de los predictores. Valores bajos de tolerancia indican que esa variable puede ser explicada por el resto de variables independientes, lo que es indicio de colinealidad.

El inverso de la tolerancia es el denominado *Factor de inflación de la varianza* (FIV), y recibe tal nombre porque se utilizan en el cálculo de las varianzas de los coeficientes de regresión. Cuanto mayor es FIV de una variable mayor es la varianza del correspondiente coeficiente de regresión. De ahí que la presencia de colinealidad conduzca a una FIV grande y por tanto a varianzas altas de los coeficientes de regresión parcial, es decir, coeficientes que pueden fluctuar mucho.

Además de los criterios de tolerancia y FIV, SPSS proporciona un diagnóstico adicional de colinealidad, aplicando un análisis de componentes principales a la matriz estandarizada no centrada de productos cruzados de las variables predictoras. La tabla con el resultado, muestra los *autovalores*, que indican el número de factores diferentes que pueden extraerse del conjunto de predictores utilizados en la ecuación de regresión. Autovalores

próximos a cero indican variables independientes muy correlacionadas entre si.

Otra información es el denominado *índice de condición* que resulta de extraer la raíz cuadrada del cociente entre el autovalor más grande y cada uno del resto de autovalores (por ejemplo el índice de condición de la dimensión 2 sería: $\sqrt{2{,}955 / 0{,}029} = 10{,}075$. En condiciones de no colinealidad, estos índices no deben superar el valor 15, y por encima de 30 indica una intensa colinealidad.

Por último, se muestran las proporciones de varianza de cada coeficiente de regresión parcial que es explicada por cada dimensión a factor. En condiciones de no colinealidad, cada dimensión suele explicar gran cantidad de varianza de un solo coeficiente (excepto la constante que aparece asociado a alguno de los otros coeficientes. Como criterio general, la colinealidad puede ser problemática cuando una dimensión con un alto índice de condición, explica gran cantidad de varianza de los coeficientes de dos o más variables.

Hay ciertos remedios que pueden resolver los problemas de colinealidad: Algunos de ellos son: el aumento del tamaño muestral; combinar en una sola variable aquellas variables que estén más relacionadas entre sí; reducir el número de variable a un grupo de factores y aplicar el análisis de regresión a las puntuaciones de esos factores; elegir de entre las variables que resultan redundantes aquellas que teóricamente sean más relevantes; aplicar la técnica de regresión *ridge*.

14.6.4.2 Casos influyentes

Se denominan así los casos que afectan notablemente al valor de la ecuación de regresión. Esto no significa que sea preciso eliminar estos casos, pero sí es conveniente disponer de medios para detectarlos, de modo que el analista pueda tomar alguna decisión racional sobre ellos. El procedimiento **Regresión** permite identificar estos casos, marcando para ello las opciones del recuadro **Distancias y Estadísticos de influencia** del cuadro de diálogo **Guardar**, tal como se muestra en la Figura 14.12.

Tres son las Distancias que expresan el grado en que cada caso se aleja del resto:

- **Mahalanobis**. Mide el grado de cada caso respecto de los promedios del conjunto de variables independientes. En regresión múltiple está distancia se obtiene multiplicando por *n-1* el valor de influencia de cada caso.
- **Cook**. Mide el cambio que se produce en las estimaciones de los coeficientes de regresión al ir eliminando cada caso de la ecuación de regresión. Distancias grandes indican un fuerte peso de dicho caso en la estimación de los coeficientes. Para evaluar las distancias puede emplearse la distribución *F* con *p+1* y *n-p-1* grados de libertad (*p* es el número de predictores y *n* el tamaño muestral). Casos con distancia de Cook superior a 1 deben ser revisados.

- **Valores de influencia**. Es una medida normalizada de la influencia potencial de cada caso, basada en el distanciamiento al centro de su distribución. Con más de 6 variables y al menos 20 casos, un valor de influencia debe ser revisado si excede $3p/n$. Los valores de influencia tienen un máximo de $(n-1)/n$. Como regla general, valores menores de 0,2 no presentan problemas; entre 0,2 y 0,5 son arriesgados; y por encima de 0,5 deben evitarse.

Respecto de los Estadísticos de influencia, se dispone de los siguientes:

- **DfBetas** (Diferencia de betas). Mide el cambio que se produce en los coeficientes de regresión al ir eliminando cada caso. SPSS crea en el *Editor de datos* tantas variables nuevas como coeficientes tiene la ecuación de regresión
- **DfBetas tipificadas**. Cociente entre DfBetas y su error típico. Un valor mayor que $2/\sqrt{n}$ delata la presencia de un posible caso de influencia. SPSS crea en el Editor de datos tantas variables nuevas como coeficientes tiene la ecuación de regresión.
- **Df Ajuste**. Mide el cambio que se produce en el pronóstico de un caso cuando ese caso es eliminado de la ecuación.
- **Df Ajuste tipificado**. Cociente entre Df Ajuste y su error típico. Se consideran casos de influencia valores por encima de $2/\sqrt{(p/n)}$.
- **Razón entre las covarianzas (RV)**. Indica en qué medida la matriz de productos cruzados cambia con la eliminación de cada caso. Un caso es de influencia, si el valor absoluto de RV-1 es mayor que $3+p/n$.

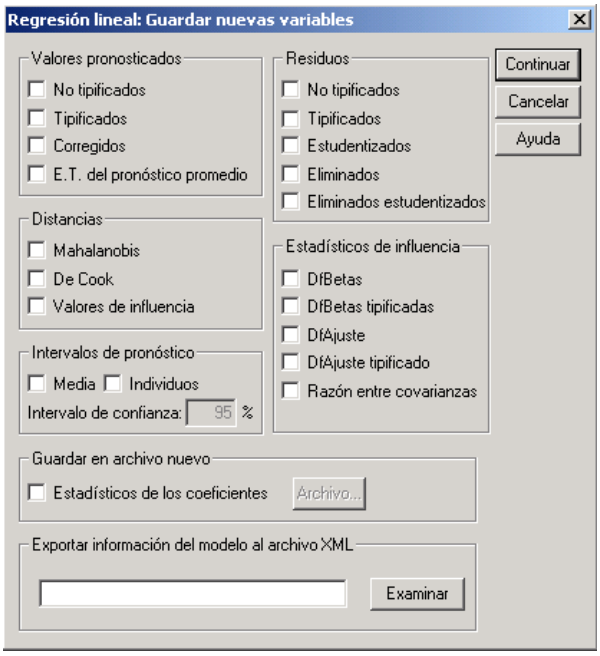


Figura 14.12. Cuadro de diálogo de Guardar nuevas variables del Regresión lineal

En la Tabla 14.9 se muestra la tabla resumen que SPSS ofrece, bajo la denominación **Estadísticos de los resíduos**. En ella se incluyen estadísticos sobre las Distancias, además de ofrecer información sobre los residuos y los pronósticos.

Tabla 14.9. Tabla con la información estadística de los residuos y de los casos de influencia

Estadísticos sobre los residuos [a]

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	4,70	13,28	9,54	1,561	90
Valor pronosticado tip.	-3,099	2,400	,000	1,000	90
Error típico del valor pronosticado	,134	,509	,220	,072	90
Valor pronosticado corregido	4,99	13,25	9,54	1,557	90
Residuo bruto	-3,20	3,19	,00	1,253	90
Residuo tip.	-2,524	2,517	,000	,989	90
Residuo estud.	-2,542	2,533	-,001	1,003	90
Residuo eliminado	-3,24	3,23	,00	1,290	90
Residuo eliminado estud.	-2,627	2,617	,000	1,016	90
Dist. de Mahalanobis	,006	13,380	1,978	2,229	90
Distancia de Cook	,000	,106	,010	,018	90
Valor de influencia centrado	,000	,150	,022	,025	90

a. Variable dependiente: Productividad

14.6.5 Métodos de obtención de la ecuación de regresión

SPSS dispone de cinco métodos para elaborar la ecuación de regresión. Estos cinco métodos se encuadran en dos grupos, según que las variables se incluyan en un solo paso o en dos pasos. Comentamos de manera resumida estos métodos, y explicamos con más detalle el criterio para seleccionar las variables que se van introduciendo en los métodos de regresión por pasos. Para seleccionar uno de los seis métodos es preciso desplegar el menú Métodos del cuadro de diálogo principal de Regresión lineal (ver Figura 14.5).

Los métodos que permiten incluir o excluir, en un solo paso, todas las variables predictoras en la ecuación son:

- **Introducir**. Se construye la ecuación de regresión introduciendo todas las variables seleccionadas en la lista Independientes. Es el método por defecto.
- **Eliminar**. Elimina en un solo paso todas las variables de la lista Independientes y ofrece los coeficientes de regresión que corresponderían a cada variable en el caso de que pasaran a formar parte de la ecuación.

Los métodos de inclusión, o exclusión, por pasos, son los siguientes

- **Hacia delante**. Las variables se incorporan una a una. La primera que se introduce es la además de superar los criterios de selección correlaciona más alto con la v. dependiente. En el paso siguiente, se introduce la variable que, además de superar los criterios de selección, presenta una correlación parcial más alta con la v. dependiente, y así sucesivamente, hasta que ya no queden variables que cumplan con el criterio de selección.

- **Hacia atrás**. Al principio se incluyen todas las variables en la ecuación, y después las va eliminando una a una. La primera que excluye es la que, además de cumplir los criterios de salida, tiene el coeficiente de regresión más bajo en valor absoluto. El proceso se detiene cuando ya no hay más variables que cumplan el criterio de salida.
- **Pasos sucesivos**. Es una mezcla de los dos métodos anteriores. Comienza, como el método *hacia delante*, seleccionando el predictor que, además de superar los criterios de entrada, correlaciona más alto con la v. dependiente. En el siguiente paso, selecciona el predictor que, además de superar los criterios de *entrada*, presenta una correlación parcial más alta con la v. dependiente. Cada vez que se incorpora un nuevo predictor, este método evalúa de nuevo la variables que ya están incorporadas para determinar si cumplen con los criterios de *salida* de la ecuación, en cuyo caso la saca del modelo. el proceso se detiene, cuando ya no hay variables que cumplan los criterios de *entrada*, ni variables en el modelo que cumplen el criterio de *salida*.

14.6.5.1 Criterios de selección/ exclusión de variables

Son varios los criterios que permiten determinar qué variables van a ser seleccionadas para formar parte de la ecuación de regresión. Entre los más generales se encuentran el valor del coeficiente de determinación múltiple, el valor del coeficiente de correlación parcial entre cada predictor y la v. dependiente, la reducción que se produce en el error típico de los residuos al incorporar un nuevo predictor, etc. El objetivo de todos estos criterios siempre es el mismo: maximizar la proporción de varianza explicada de la variable dependiente, utilizando el mínimo de predictores posible.

En los tres métodos de selección por pasos hay dos criterios de tipo estadístico:

- **Criterio de significación**. Según este criterio sólo se incorporar al modelo de regresión las variables que contribuyen significativamente a su ajuste. La contribución de cada variable se establece contrastando, mediante el coeficiente de correlación parcial, la hipótesis de independencia entre ese predictor y la v. dependiente. Para tomar una decisión se siguen dos criterios (cuyos valores puede modificar el analista):
  - **Probabilidad de F**. Un predictor formará parte del modelo si el nivel crítico asociado a su coeficiente de correlación parcial al contrastar la hipótesis de independencia es menor que 0,05 (probabilidad de entrada). Y queda fuera del modelo si el nivel crítico es mayor que 0,10 (criterio de salida).
  - **Valor de F**. Un predictor se incorpora al modelo si el valor del estadístico F utilizado para contrastar la hipótesis de independencia es mayor que 3,84 (valor de *entrada* por defecto). Y saldrá del modelo si el valor de F es menor que 2,71 (valor de *salida* por defecto)

  Para modificar estos valores pulsar el botón Opciones del cuadro de diálogo Regresión lineal, y se muestra el cuadro de la Figura 14.13.

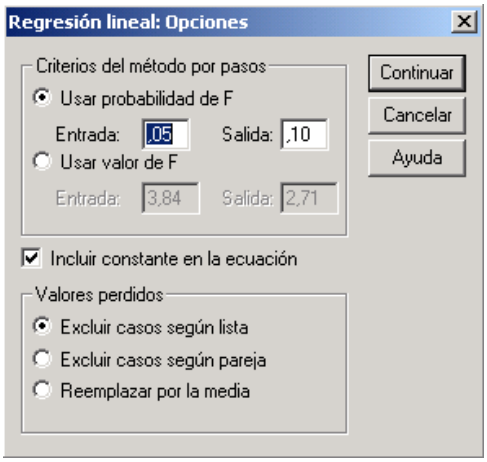


Figura 14.13 Cuadro de Opciones de Regresión lineal

- **Criterio de tolerancia**. Si se supera el nivel de significación, un predictor sólo pasa a formar parte del modelo si su nivel de tolerancia es mayor que el nivel establecido por defecto (0,0001), valor que puede ser cambiado mediante sintaxis, y si, además, incluso correspondiéndole un coeficiente de correlación parcial significativamente distinto de cero, su incorporación al modelo hace que alguna de los predictores seleccionadas previamente pase a tener un nivel de tolerancia por debajo de nivel establecido por defecto.

Una información que aporta SPSS (si se marca la opción en el cuadro de diálogo **Estadísticos**), es el cambio que se produce en R^2 a medida que se van incorporando (o excluyendo) variables al modelo. El cambio se define como $R^2_{cambio} = R^2 - R^2_i$, donde R^2_i se refiere al coeficiente de determinación obtenido con todas las variables excepto la *i*-ésima . Un cambio grande en R^2 indica que esa variable contribuye significativamente a explicar la v. dependiente. El estadístico F que contrasta si R^2_{cambio} es significativamente distinto de cero es:

$$F_{cambio} = \frac{R^2_{cambio} / q}{(1 - R^2) / (N - p - 1)} = \frac{R^2_{cambio}(N - p - 1)}{q(1 - R^2)}$$

que sigue la distribución F con *q* y (*N-p-1*) grados de libertad, siendo *N* el número total de casos; *p* el número de predictores en la ecuación; y *q* el número de variables predictoras que entran que entran en ese paso. En la Tabla 14.10 se ven los valores de R^2_{cambio} y su significación estadística cuando utilizamos el método paso a paso en las variables que están sirviendo de ejemplo.

Tabla 14.10 Tablas con las *Variables ingresadas* en cada paso y *Resumen del modelo* con los cambios en R^2

Variables introducidas/eliminadas [a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Atención para el detalle	,	Por pasos (criterio: Prob. de F para entrar <= ,050, Prob. de F para salir >= ,100).
2	Destreza manual	,	Por pasos (criterio: Prob. de F para entrar <= ,050, Prob. de F para salir >= ,100).

a. Variable dependiente: Productividad

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,671[a]	,451	,444	1,492	,451	72,179	1	88	,000
2	,780[b]	,608	,599	1,267	,157	34,917	1	87	,000

a. Variables predictoras: (Constante), Atención para el detalle

b. Variables predictoras: (Constante), Atención para el detalle, Destreza manual

El lector puede comprobar los valores de *Cambio en F* en cada uno de los dos modelos. En el primero, entra la variable "Atención para el detalle" con un R^2 de 0,4506 (en el redondeo queda 0,451), por tanto el valor de F es:

$$F_{cambio} = \frac{0{,}4506(100-1-1)}{1(1-0{,}4506)} = 72{,}179$$

Cuando en el segundo paso entra el otro predictor (Destreza), R^2 aumenta hasta 0,6079 (redondeado a 0,608) lo que supone un cambio de 0,1573 (redondeado a 0,157) (0,6079 – 0,4506 = 0,1573). La F de este cambio es:

$$F_{cambio} = \frac{0{,}1573(100-2-1)}{1(1-0{,}6079)} = 34{,}9170$$

En este caso las dos variables predictoras contribuyen significativamente a explicar la variabilidad de la v. independiente.

14.6.5.2 Variables que debe incluir un modelo de regresión

La ventaja del método por pasos para elaborar la ecuación de regresión, es que podemos construir un modelo parsimonioso, en el sentido de que un mínimo de predictores pueden explicar el máximo de la varianza. Es verdad que a medida que incorporamos predictores en el modelo R^2 permanece igual o va aumentando (nunca disminuye), pero no siempre la cuantía de este aumento justifica la inclusión de nuevas variables. Por regla general, ha medida que el ajuste del modelo es mayor, la estimaciones que hace el modelo son más preciso (tienen menos error), pero no siempre el incremento en R^2 se corresponde con una disminución de los residuos.

Con cada variable nueva la suma de cuadrados de la regresión aumenta un grado de libertad y la suma de cuadrados de los residuos lo pierde, lo que conlleva un aumento de la *media cuadrática residual*. De donde se deduce que es preciso mantener un equilibrio entre lo que se gana en potencia explicativa y el aumento que se da en el *error típico residual*. Si la ganancia en la explicación de la variabilidad compensa la pérdida de grados de libertad en los residuos, sí se justifica la inclusión de una nueva variable aunque el aumento de R^2 no sea muy grande.

Otro elemento que deberemos considerar es el del tamaño muestral. Como el lector sabe, a medida que aumenta la muestra aumenta también la potencia de los contrastes estadísticos, lo que puede conducir a que efectos muy pequeños resulten ser significativos estadísticamente. Por el contrario, con muestras pequeñas, el efecto diferencial debe ser muy grande para que resulte estadísticamente significativo, y dicha significación estadística suele coincidir con que sea importante el efecto en el ámbito teórico. Por tanto, el tamaño de la muestra también es un criterio a tener en cuenta: si es muy grande, es conveniente tomar en consideración otros elementos, además del criterio de significación estadística.

Por último, es preciso señalar que al aumentar el número de variables predictores, el puro azar puede conducir a falsos positivos (error tipo I). Un modo de evitar este inconveniente es, con una muestra grande, dividirla en dos, construir un modelo con una de las submuestras y verificar el resultado en la otra submuestra.

14.6.6 Pronósticos generados en el procedimiento Regresión lineal

Anteriormente se ha visto que SPSS genera una serie de variables que se almacenan en el archivo de trabajo. Estas variables se marcan en el cuadro de diálogo al que se accede (ver Figura 14.12) pulsando el botón **Guardar** del cuadro de diálogo principal de **Regresión lineal**. De entre todas las nuevas variables que se muestran en el cuadro de diálogo, en este apartado hacemos referencia sólo a los diferentes pronósticos que se generan. Las nuevas variables de pronósticos reciben un nombre seguido de un número de serie: *nombre_#*. Si en posteriores análisis se solicitan estos mismos pronósticos, se genera una nueva variable con el mismo nombre y un valor más de número de serie, y así sucesivamente. Los diferentes pronósticos que se pueden obtener son:

- **No tipificados** (nombre: *pre_#*). Pronósticos en puntuaciones directas.

- **Tipificados** (nombre: *zpr_#*). Pronósticos tipificados (cada pronóstico no tipificado se resta a la media y se divide por la desviación típica).
- **Corregidos** (nombre: *adj_#*). Pronóstico que corresponde a cada caso cuando la ecuación de regresión se calcula sin incluir ese caso.
- **E.T. del pronóstico promedio** (nombre: *sep_#*).. Error típico de los pronósticos correspondientes a los casos que tienen el mismo valor en las variables independientes. Al efectuar pronósticos es posible optar entre dos tipos: 1) pronóstico individual para cada valor X_i, y 2) pronosticar para cada caso el pronóstico promedio correspondiente a todos los casos con el mismo valor X_i. En ambos casos se obtiene el mismo pronóstico, pero el error típico del pronóstico individual es siempre igual o mayor que el del pronóstico promedio.

Además de estos pronósticos y el error típico se pueden obtener sus intervalos de confianza:

- **Media**. Intervalo de confianza basado en los errores típicos de los pronósticos promedio
- **Individuos**. Intervalo de confianza basado en los errores típicos de los pronósticos indivuales.

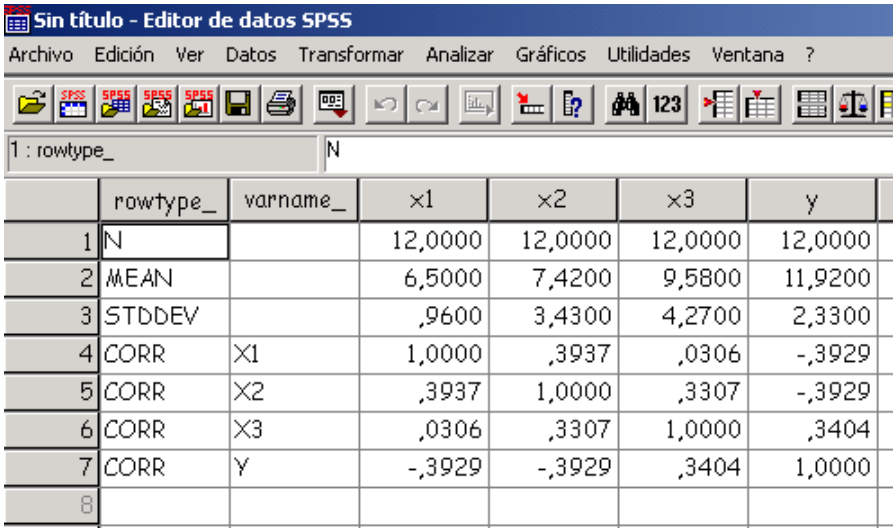
14.6.7 Regresión múltiple a partir de una matriz de correlaciones

SPSS permite obtener, mediante sintaxis, un modelo de regresión a partir de la matriz de correlaciones de una serie de variables. Lógicamente, al no disponer de casos no es posible realizar el contraste de los supuestos básicos del modelo de regresión, que como el lector ha visto, se basan en los residuos, por lo cual se tiene que suponer que estos supuestos se cumplen. Tampoco se pueden realizar el análisis de los casos influyentes, ni generar gráficos de dispersión.

Este tipo de datos en forma de matriz se puede escribir mediante sintaxis, o directamente en el **Editor de datos**. Para escribirlos en una ventana de sintaxis el programa quedaría de la siguiente forma:

```
MATRIX DATA VARIABLES= X1 X2 X3 Y
/ CONTENTS= MEAN SD N CORR / FORMAT= UPPER
NODIAGONAL.
BEGIN DATA
6,50  7,42  9,58  11,92
0,96  3,43  4,27  2,33
12  12  12  12
0,3937  0,0306  -0,3929
0,3307  -0,3929
0,3404
END DATA.
```

En la primera fila se especifica que los datos posteriores tienen forma de matriz (MATRIZ DATA). Después de especificar el nombre de las variables tras el subcomando VARIABLES, se señala el contenido (CONTENTS) de las filas de datos (Media –MEAN-; desviación típica –SD-; número de casos –N-; y correlaciones –CORR-; a continuación se especifica el formato (FORMAT) de la matriz de correlaciones: en este caso UPPER NODIAGONAL, y que significa que la matriz es triangular superior y no contiene la diagonal con los 1,000 (correlación de cada variable consigo misma). A continuación, de BEGIN DATA, se escriben los estadísticos en el mismo orden que se han especificado en CONTENTS. Después de escribir los datos se termina con END DATA. Una vez ejecutado estas líneas de sintaxis, los datos en el Editor de Datos quedan de la manera que se puede ver en la Figura 14.14.



	rowtype_	varname_	x1	x2	x3	y
1	N		12,0000	12,0000	12,0000	12,0000
2	MEAN		6,5000	7,4200	9,5800	11,9200
3	STDDEV		,9600	3,4300	4,2700	2,3300
4	CORR	X1	1,0000	,3937	,0306	-,3929
5	CORR	X2	,3937	1,0000	,3307	-,3929
6	CORR	X3	,0306	,3307	1,0000	,3404
7	CORR	Y	-,3929	-,3929	,3404	1,0000
8						

Figura 14.14. Datos en forma de matriz de correlaciones en el Editor de datos

En el Editor de datos, además de las variables nombradas por el usuario, se generan dos variables: *rowtype_* y *varname_*. La primera, *rowtype_*, es la de contenidos; la segunda, *varname_*, contiene el nombre de las variables con las que vamos a realizar el análisis. Con este tipo de datos, el análisis no se puede efectuar a través de menús, sino sólo mediante sintaxis. El procedimiento básico con los datos que se muestran en la Figura 14.14 se puede escribir de la siguiente forma:

```
REGRESSION MATRIX=IN(*) /VARIABLES=X1 TO Y
/DEP=Y
/ENTER.
```

y las tablas con los resultados que se muestran en el Visor son las siguientes:

Variables introducidas/eliminadas [b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	X3, X1, X2 [a]	,	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: Y

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,666[a]	,444	,236	2,0371326

a. Variables predictoras: (Constante), X3, X1, X2

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	26,519	3	8,840	2,130	,175[a]
	Residual	33,199	8	4,150		
	Total	59,718	11			

a. Variables predictoras: (Constante), X3, X1, X2

b. Variable dependiente: Y

Coeficientes [a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	15,179	4,459		3,404	,009
	X1	-,540	,701	-,223	-,771	,463
	X2	-,320	,208	-,472	-1,543	,161
	X3	,275	,153	,503	1,790	,111

a. Variable dependiente: Y

Este tipo de datos en forma de matriz de correlaciones, con los estadísticos media, desviación típica y número de casos, también se pueden escribir directamente en el **Editor de datos**, respetando el nombre de las variables que genera SPSS cuando se escriben por medio de la sintaxis, así como el nombre de los contenidos.

Para terminar este capítulo, vamos a proponer un ejercicio al lector: un análisis con este tipo de datos en forma de matriz. Los datos se pueden ver en la Figura 14.15, y están tomados del libro "*Psicometría: teoría de los tests psicológicos y educativos*", página 455, de Rosario Martínez Arias, editado por Síntesis en el año 1996.

	rowtype_	varname_	curso	estim	conoc	interés	claridad	evalpr
1	N		30,0000	30,0000	30,0000	30,0000	30,0000	30,0000
2	MEAN		31,2000	32,5667	18,3000	18,7667	28,8667	25,5667
3	STDDEV		11,2661	10,2774	7,1301	7,7579	8,9587	8,2656
4	CORR	CURSO	1,0000	,5100	,0370	,0100	,3660	,5150
5	CORR	ESTIM	,5100	1,0000	,1750	,1680	,5870	,6920
6	CORR	CONOC	,0370	,1750	1,0000	,2760	,2790	,3910
7	CORR	INTERÉS	,0100	,1680	,2760	1,0000	,3580	,4950
8	CORR	CLARIDAD	,3660	,5870	,2790	,3580	1,0000	,7890
9	CORR	EVALPR	,5150	,6920	,3910	,4950	,7890	1,0000

Figura 14.15. Datos página 455 del libro: "Psicometría: teoría de los tests psicológicos y educativos"

Las variables son los datos recogidos en 6 cuestionarios que cumplimentan los estudiantes de un colegio. El nombre de las variables significa lo siguiente:

CURSO. Evaluación de la materia impartida.

ESTIM. Estimulación proporcionada por el profesor.

CONOC. Conocimientos de la materia.

INTERÉS. Interés del profesor por la materia.

CLAR. Claridad en la exposición.

EVALPR. Evaluación del profesor.

Se trata de construir un modelo que permita predecir la Evaluación que hacen los alumnos del profesor (EVALPR), tomando como predictores los otros cinco cuestionarios. En primer lugar se va a elaborar un modelo en el que entren todas las predictoras, y la sintaxis para ello es la siguiente:

```
REGRESSION MATRIX = IN (*)
  /VARIABLES = curso TO evalpr
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHANGE ZPP
  /CRITERIA= PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT evalpr
  /METHOD= ENTER curso estim conoc interés claridad .
```

Para que el lector coteje si sus resultados son correctos el resumen del modelo es el que se muestra a continuación:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,901[a]	,812	,773	3,9350823	,812	20,790	5	24	,000

a. Variables predictoras: (Constante), Claridad de la exposición, Conocimientos de la materia, Evaluación materia impartida, Interés del profesor por la materia, Estimulación proporcionada por el profesor

Para el siguiente análisis se utiliza el método *Por pasos sucesivos* (*Stepwise*), y la sintaxis es la siguiente:

```
REGRESSION MATRIX = IN (*)
  /VARIABLES = curso TO evalpr
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT evalpr
  /METHOD=STEPWISE curso estim conoc interés claridad  .
```

y la tabla con el resumen del modelo es:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,789[a]	,623	,609	5,1682173	,623	46,176	1	28	,000
2	,838[b]	,702	,680	4,6728873	,080	7,251	1	27	,012
3	,873[c]	,762	,734	4,2600041	,059	6,487	1	26	,017

a. Variables predictoras: (Constante), Claridad de la exposición

b. Variables predictoras: (Constante), Claridad de la exposición, Estimulación proporcionada por el profesor

c. Variables predictoras: (Constante), Claridad de la exposición, Estimulación proporcionada por el profesor, Interés del profesor por la materia

15. Pruebas no paramétricas

15.1 Introducción

Los procedimientos de contraste de hipótesis (pruebas T, ANOVA, análisis de regresión, prueba de homogeneidad de varianzas, etc.) que se han estudiado en los capítulos precedentes, tenían en común tres aspectos:

- Contrastan hipótesis sobre parámetros (medias, varianzas, coeficientes de correlación, coeficientes de regresión, etc.).
- Se tienen que cumplir una serie de supuestos para su correcta aplicación (independencia, normalidad y homocedasticidad).
- El nivel de medida de los datos tiene que ser de escala (intervalo o razón).

En la literatura estadística este tipo de contrastes se denominan paramétricos, y los estadísticos de estos contrastes se distribuyen de acuerdo a tres modelos de probabilidad: normal, *t* de Student y *F* de Snedecor-Fisher. Dado que los supuestos para su aplicación son bastante exigentes, este tipo de contrastes es muy común en el ámbito de las ciencias que trabajan con variables cuyo nivel de medida es siempre de escala. Sin embargo, en el ámbito de las ciencias sociales, los fenómenos objeto de estudio no siempre son factibles de cuantificar con un nivel de medida del tipo que requiere estos contrastes. En realidad, en el ámbito concreto de la Psicología lo más frecuente es trabajar con escalas de tipo ordinal o nominal, y las poblaciones sobre las que se trabaja, no siempre cumple que se distribuya aproximadamente normal.

Con el objetivo de disponer de alternativas a estos contrastes, los estadísticos, a lo largo del último siglo, han ido diseñando instrumentos de análisis para realizar pruebas estadísticas que no precisan del cumplimiento de los supuestos de los contrastes paramétricos. Algunos permiten realizar contrates sobre la tendencia central de la población –en general, sobre índices de posición– sin necesidad de que el nivel de medida de la variable sea de escala; otros permiten realizar contrastes sobre la forma de la distribución. A todos ellos se les engloba bajo la denominación de *contrastes no paramétricos* (*pruebas no paramétricas*, en la terminología de SPSS).

Todas la pruebas no paramétricas disponibles en SPSS se encuentran en la opción **Pruebas no paramétricas** del menú **Analizar** y su secuencia es la siguiente:

- **Pruebas para una muestra**: Chi-cuadrado, Binomial, Rachas y Kolmogorv–Smirnov
- **Pruebas para dos muestras independientes**: U de Mann–Whitney, Kolmogorov–Smirnov, Reacciones extremas de Moses y Rachas de Wald–Wolfowitz.
- **Pruebas para varias muestras independientes**: H de Kruskal–Wallis y Mediana.

- **Pruebas para dos muestras relacionadas**: Wilcoxon, Signos y McNemar.
- **Pruebas para varias muestras relacionadas**: Friedman, W de Kendall y Q de Cochran

(**Nota**: Los datos que vamos a utilizar para ilustrar estas pruebas son los del archivo *Datos de empleados*, que se encuentra en el CD del programa).

15.2 Pruebas para una muestra

Dentro de las pruebas para una muestra, se puede establecer la bondad de ajuste tanto con variable categóricas (*Chi-cuadrado*) como con variables de escala (*Kolmogorv–Smirnov*). También se puede establecer un contraste sobre proporciones y cuantiles (*Binomial*) y determinar si una muestra de datos ha sido extraída de forma aleatoria (*Rachas*).

15.2.1 Pruebas Chi–cuadrado

Esta prueba permite determinar si una distribución empírica se ajusta o no a determinada distribución teórica, bien sea binomial, uniforme, mutinomial. El estadístico de contraste, propuesto por Pearson compara las frecuencias observadas con las frecuencias esperadas si los datos tuvieran una distribución determinada. En caso de que la hipótesis nula fuera correcta el nivel crítico del valor del estadístico debería ser por encima de 0,05. Su formula es la siguiente:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - \hat{n}_i\right)^2}{\hat{n}_i}$$

donde:

n_i es la frecuencia observado en la categoría *i-ésima*, y

$\hat{n}_i$ es la frecuencia esperada o teórica en la categoría *i-ésima*

Las frecuencias teóricas (esperadas) se obtienen multiplicando la proporción teórica de cada categoría π_i, por el número de casos válidos; es decir: $n\pi_i$. Si no hay casillas vacias y le número de frecuencias esperadas menores de 5 no supera el 20% del total de frecuencias esperadas, este estadístico se distribuye según el modelo *chi-cuadrado* con *k–1* grados de libertad, siendo *k* el número de categorías de la variable.

El cuadro de diálogo de la prueba Chi–cuadrado es el que se muestra en la Figura 15.1.

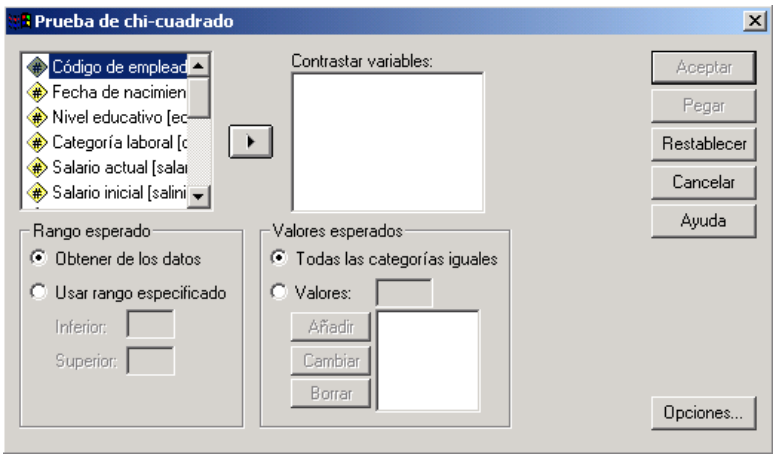


Figura 15.1. Cuadro de diálogo de la Prueba de Chi–cuadrado

Una vez trasladada la variable a la lista contrastar variables hay que definir algunos aspectos: Respecto del rango de los datos, es posible determinar el rango de valores que será objeto de análisis. Si se especifica la opción por defecto (**Obtener de los datos**), cada valor diferente se considera como una categoría. Si se especifica un rango, sólo se tendrán en cuenta en el análisis los valores incluidos dentro del rango especificado, excluyendo el resto de valores.

Respecto de los valores esperados, se pueden especificar el tipo de distribución que se va a contrastar. Si se señala la opción **Todas las categorías iguales**, la distribución de contraste será uniforme: las frecuencias esperadas se obtienen dividiendo el número total de casos válidos entre el número de categorías (o valores diferentes) de la variable. Con la opción **Valores**, se pueden especificar las frecuencias relativas (proporciones) esperadas concretas para cada una de las categorías de la variable. Así, por ejemplo, si una variable tiene tres categorías y se ingresan los valores 4, 4 y 2 (o también 40, 40 y 20), SPSS interpreta que la frecuencia relativa esperada de la primera categoría es 0,4, la de la segunda es 0,4 y la de la tercera 0,2. Es fácil colegir que este es el procedimiento para contrastar una distribución binomial o una distribución multinomial.

Veamos como ejemplo, el contraste de la variable Categoría laboral, primero como una distribución uniforme y, posteriormente como una distribución multinomial, con frecuencias relativas esperadas 0,7 para Administrativo, 0,1 para Seguridad y 0,2 para Directivo. Las tablas resultantes, se pueden ver en la Tabla 15.1.

Tabla 15.1 Tablas de resultados del contraste de Chi–cuadrado para una distribución uniforme y otra multinomial (con proporciones 0,7, 0,1 y 0,2).

Contraste de *Categoría laboral* como una distribución uniforme

Categoría laboral

	N observado	N esperado	Residual
Administrativo	363	158,0	205,0
Seguridad	27	158,0	-131,0
Directivo	84	158,0	-74,0
Total	474		

Estadísticos de contraste

	Categoría laboral
Chi-cuadrado [a]	409,253
gl	2
Sig. asintót.	,000

a. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 158,0.

Contraste de *Categoría laboral* como una distribución multinomial

Categoría laboral

	N observado	N esperado	Residual
Administrativo	363	331,8	31,2
Seguridad	27	47,4	-20,4
Directivo	84	94,8	-10,8
Total	474		

Estadísticos de contraste

	Categoría laboral
Chi-cuadrado [a]	12,944
gl	2
Sig. asintót.	,002

a. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 47,4.

Se observa que en ambos contrastes no se puede aceptar la hipótesis planteada: ni la distribución es uniforme, ni es multinomial con las proporciones-probabilidades especificadas.

15.2.2 Prueba Binomial

Esta prueba permite contrastar proporciones de variables dicotómicas o dicotomizadas. Las variables dicotómicas son aquellas que sólo presentan dos modalidades o categorías (sexo, pruebas de verdadero falso, etc.). Las variables

dicotomizadas son variables de un nivel de medidas superior al nominal sobre la que se establece un dicotomía, tomando como punto de corte algunos de los valores de la variable (por ejemplo, la dicotomización Apto – No Apto que se establece sobre las calificaciones de las asignaturas). En uno u otro caso, se puede contrastar un determinado modelo de probabilidad binomial con parámetros *n* (tamaño de la muestra) y *π* (proporción de aciertos). Cuando el tamaño de *n* no supera 25, la distribución que toma SPSS para contrastar la probabilidad asociada a cada valor de la variable X es la Binomial. Cuando es superior a 25, la distribución se aproxima a la normal con parámetros: $E(X) = n\pi$ y $\sigma_X = \sqrt{n\pi(1-\pi)}$, y la expresión del estadístico de contraste Z es:

$$Z = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$$

que se distribuye según el modelo normal N(0,1). Cuando utiliza la aproximación a la normal, SPSS realiza la denominada corrección por continuidad, que consiste en sumar (si X es menor que nπ) o restar (si X es mayor que nπ) 0,5 puntos a X para hacer algo más conservador el contraste. De este modo, el estadístico de contraste será el siguiente:

$$Z = \frac{X \pm 0{,}5 - n\pi}{\sqrt{n\pi(1-\pi)}}$$

El cuadro de diálogo de la prueba binomial, es el que se muestra en la Figura 15.2.

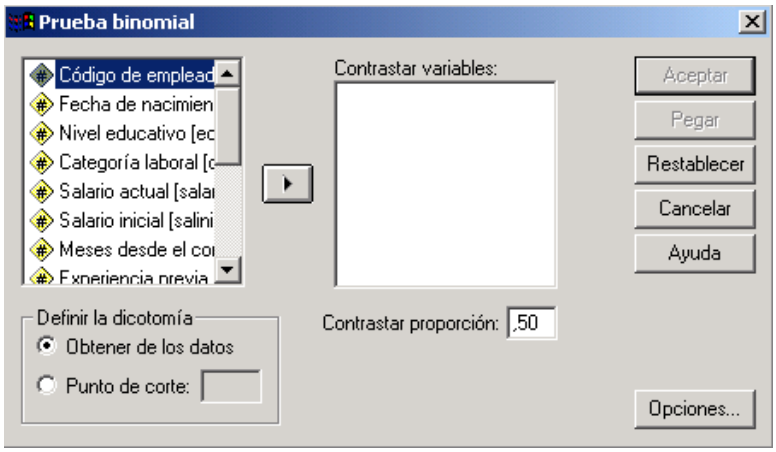


Figura 15.2 Cuadro de diálogo de la Prueba binomial

Por defecto, la dicotomía se obtiene de los propios datos, y la proporción de contraste e 0,5. No obstante, como ya se ha comentado, se pueden realizar la prueba sobre variable cuantitativas u ordinales, las cuales se dicotomizan señalando un punto que sirva para cortar la distribución en dos categorías. Esta opción útil para contrastar hipótesis sobre percentiles en la población. Esta prueba se conoce en la literatura estadística como la *prueba de los signos*.

Para ilustrar el procedimiento de la prueba *binomial*, en primer lugar, vamos a contrastar la hipótesis de que la población de empleados está repartida al 50% entre hombres y mujeres. Para ello, primero se transforma la variable sexo que es de cadena en una variable numérica, mediante el procedimiento Remodificación

automática. Una vez generada esa nueva variable, ya se puede borrar la antiguo y renombrar la nueva con el mismo nombre (sexo) de la antigua. La razón para esta remodificación, es que la prueba binomial, sólo se aplica a variables de escala. Después se va a contrastar la hipótesis de que el percentil 75 de años de escolarización (variable *educ*) es 14. Las tablas de resultados para estos dos contrastes se muestran en la Tabla 15.2.

Tabla 15.2. Resultados de la prueba binomial para una variable dicotómica con parámetro p = 0,5, y para una variable de escala dicotomizada con parámetro p = 0,75

Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (bilateral)
Sexo	Grupo 1	Hombre	258	,54	,50	,060[a]
	Grupo 2	Mujer	216	,46		
	Total		474	1,00		

a. Basado en la aproximación Z.

Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (unilateral)
Nivel educativo	Grupo 1	<= 16	424	,89	,75	,000[a]
	Grupo 2	> 16	50	,11		
	Total		474	1,00		

a. Basado en la aproximación Z.

Prueba binomial (sobre una muestra aleatoria de 19 casos)

		Categoría	N	Proporción observada	Prop. de prueba	Sig. exacta (bilateral)
Sexo	Grupo 1	Mujer	9	,47	,50	1,000
	Grupo 2	Hombre	10	,53		
	Total		19	1,00		

Se observa que cuando la proporción de contraste es 0,5 el contraste es bilateral, mientras que cuando es mayor o menor de ese valor el contraste es unilateral. En el caso del contraste *bilateral*, el nivel crítico se obtiene multiplicando por 2 la probabilidad de encontrar un número de casos igual o mayor que el de la categoría de referencia (si la proporción de casos de dicha categoría es mayor que 0,5) o multiplicando por dos la probabilidad de encontrar un número de casos igual o menor que el de la categoría de referencia (si la proporción de casos de dicha categoría es menor que 0,5). Cuando el contraste es *unilateral*, el nivel crítico resulta ser la probabilidad de encontrar un número de casos igual o mayor, o igual o menor, según sea la proporción de casos de la categoría de referencia.

La distribución de referencia para este ejemplo es la normal, ya que el tamaño de la muestra es superior a 25, y así se señala en la nota al pie de tabla. Si la misma prueba la realizamos sobre una muestra menor a igual de 25 la distribución de referencia sería la binomial, que es la distribución de referencia por defecto y no

se muestra ninguna nota a pie de tabla, tal como se puede ver en la tercera tabla de la Tabla 15.2.

15.2.3 Prueba de *rachas*

Esta prueba sirve para contrastar si una muestra de observaciones ha sido extraída o no de forma aleatoria, es decir, independientes entre sí. El concepto de racha se refiere a la secuencia de observaciones de un mismo tipo. Si, por ejemplo, tenemos un grupo de personas, (**H**)ombres y (**M**)ujeres, para elegir una muestra de 10, y se obtiene la siguiente secuencia:

HHMMMMHMHH

en dicha muestra habrá exactamente 5 rachas: HH–MMMM–H–M–HH. En total hay en la muestra la misma proporción de hombres y mujeres. Si en lugar de este resultado se hubiera obtenido este otro:

MMMMMHHHHH

también habría el mismo número de hombres y mujeres pero la secuencia de extracción no parece que presente el mismo carácter aleatorio que la secuencia anterior.

La prueba de rachas permite determinar si el número observado de éstas (R) en una muestra de tamaño n es lo suficientemente grande o lo suficientemente pequeña como para poder rechazar la hipótesis de que la muestra se ha extraído de manera aleatoria, y por tanto las observaciones son independientes.

Para obtener el número de rachas es necesario que las observaciones estén clasificadas en dos grupos exhaustivos y excluyentes, o si no lo están, se debe utilizar algún criterio (media, mediana, moda, o algún otro valor) para dicotomizar la muestra. Una vez clasificadas las observaciones en dos grupos de tamaños n_1 y n_2, SPSS utiliza una tipificación del número de rachas para realizar el contraste:

$$Z = \frac{R - E(R)}{\sigma_R}$$

siendo $E(R) = {(2n_1n_2)}/{(n+1)}$ y $\sigma_R = \sqrt{{[2n_1n_2(2n_1n_2 - n)]}/{[n^2(n-1)]}}$.

Este estadístico Z tiene una distribución normal N(0,1). SPSS ofrece el nivel crítico bilateral que resulta de multiplicar por 2 la probabilidad de encontrar un número de rachas igual o menor que el encontrado (R < E(R)) o igual o mayor que el encontrado (R > E(R)). Si el tamaño de la muestra es menor de 50, el estadístico Z se obtiene realizando la corrección por continuidad de acuerdo a los siguientes criterios:

Si R – E(R) < – 0,5, se suma 0,5 a R, con lo que $Z = [R + 0{,}5 - E(R)]/\sigma_R$.

Si R – E(R) > 0,5, se resta 0,5 a R, con lo que $Z = [R - 0{,}5 - E(R)]/\sigma_R$.

Si –0,5 ≥ [R – E(R)] ≤ 0,5, entonces Z = 0.

El cuadro de diálogo de esta prueba es el que se muestra en la Figura 15.3.

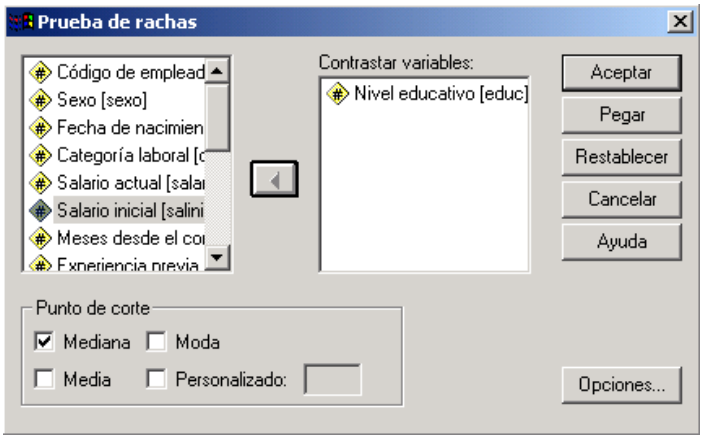


Figura 15.3 Cuadro de diálogo de la *Prueba de rachas*

Para que la prueba de rachas tenga utilidad estadística, es preciso aplicarla a los datos como originalmente se han recogido, sin que medie proceso alguno de ordenación de la variable estudiada, ya que en este caso los valores por debajo del punto de corte estarán a un lado y los valores iguales o por encima del punto de corte estarán a otro y en consecuencia sólo habrá dos rachas. Si realizamos la *prueba de rachas* a la variable nivel educativo (*educ*), estando el archivo *Datos de empleados* ordenado por el código de empleado (variable *id*), el resultado de la prueba es el que se muestra en la Tabla 15.3

Tabla 15.3 Resultados de la *prueba de rachas* para la variable *educ*, tomando como punto de corte la mediana

Prueba de rachas

	Nivel educativo
Valor de prueba [a]	12,00
Casos < Valor de prueba	53
Casos >= Valor de prueba	421
Casos en total	474
Número de rachas	97
Z	,430
Sig. asintót. (bilateral)	,667

a. Mediana

Se observa en la tabla el criterio que sigue SPSS para clasificar respecto del punto de corte. Los valores que están por debajo del valor de prueba (punto de corte) pertenecen a una categoría y los son iguales o mayores del punto de corte pertenecen a otra categoría. En este caso el nivel crítico nos indica que no podemos rechazar la hipótesis de independencia entre las observaciones.

15.2.4 Prueba de Kolmogorov–Smirnov (K–S) para una muestra

Con esta prueba se pueden realizar contrastes para determinar si una muestra se *ajusta* a una distribución teórica de probabilidad. En este sentido es similar a la prueba chi–cuadrado, sólo que en esta prueba se evalúa la distribución de variables de escala. La prueba compara dos funciones de distribución, la empírica $F(X_i)$ de los

datos y la teórica de contraste $F_0(X_i)$. Son cuatro las distribuciones que se pueden contrastar: *Normal*, *Uniforme*, *Poisson* y *Exponencial*.

La función de distribución empírica se obtiene ordenando ascendentemente los valores X_i, y posteriormente se calcula la proporción acumulada de cada valor: i/n, donde i es el rango de cada observación. Posteriormente esta función de distribución empírica se compara con la función de distribución teórica. El estadístico de K–S se calcula a partir de la diferencia D_i más grande existente entre $F(X_i)$ y $F_0(X_i)$:

$$Z_{K-S} = \text{máx}|D_i|\sqrt{n}$$

que sigue la distribución normal N(0,1).

El cuadro de diálogo de este procedimiento es el que se muestra en la Figura 15.4.

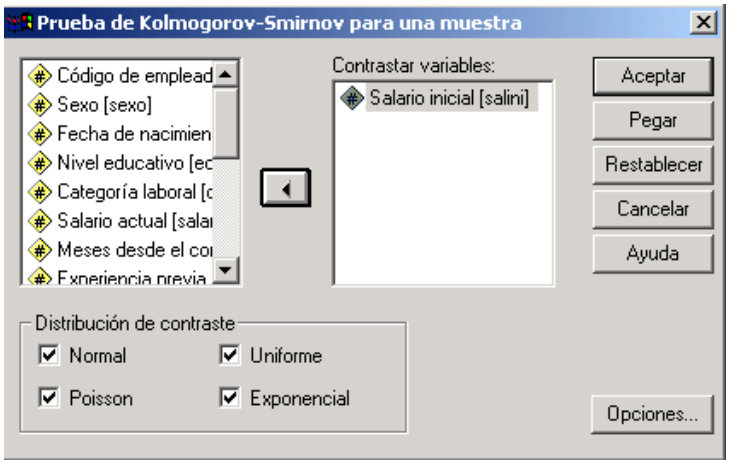


Figura 15.4 Cuadro de diálogo del procedimiento Prueba de Kolmogorov-Smirnov

Como ejemplo contrastamos la distribución de la variable salario inicial con las cuatro distribuciones teóricas de referencia, y los resultados pueden verse en la Tabla 15.4

Tabla 15.4 Resultados de la *prueba de Kolmogorov-Smirnov* para la variable *salario inicial*, contrastada con las cuatro distribuciones teóricas de referencia

Prueba de Kolmogorov-Smirnov para una muestra

		Salario inicial
N		474
Parámetros normales [a,b]	Media	$17,016.09
	Desviación típica	$7,870.638
Diferencias más extremas	Absoluta	,252
	Positiva	,252
	Negativa	-,170
Z de Kolmogorov-Smirnov		5,484
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Prueba de Kolmogorov-Smirnov para una muestra 2

		Salario inicial
N		474
Parámetros uniformes [a,b]	Mínimo	$9,000
	Máximo	$79,980
Diferencias más extremas	Absoluta	,677
	Positiva	,677
	Negativa	-,002
Z de Kolmogorov-Smirnov		14,737
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es la Uniforme.

b. Se han calculado a partir de los datos.

Prueba de Kolmogorov-Smirnov para una muestra 3

		Salario inicial
N		474
Parámetro de Poisson [a,b]	Media	$17,016.09
Diferencias más extremas	Absoluta	,726
	Positiva	,726
	Negativa	-,253
Z de Kolmogorov-Smirnov		15,800
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es la de Poisson.

b. Se han calculado a partir de los datos.

Prueba de Kolmogorov-Smirnov para una muestra 4

		Salario inicial
N		474
Parámetro exponencial. [a,b]	Media	$17,016.09
Diferencias más extremas	Absoluta	,428
	Positiva	,136
	Negativa	-,428
Z de Kolmogorov-Smirnov		9,312
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es exponencial.

b. Se han calculado a partir de los datos.

La estructura de la tabla con el resultado del contraste es siempre la misma: en el apartado Diferencias más extremas se reflejan la más alta en valor absoluto y las más altas en sentido positivo y negativo, después se refleja el valor del estadístico Z (por ejemplo, para el ajuste de los datos a una distribución normal, $Z = 5{,}484 = 0{,}252 \times \sqrt{474}$), y por último el nivel crítico del valor de Z (contraste bilateral).

15.3 Prueba para dos muestras independientes

Con estas pruebas permiten realizar contrastes sobre datos que contienen una variable independiente categórica (definida siempre como variable numérica) sobre la que se pueden definir 2 grupos y una variable dependiente medida al menos a novel ordinal. Se trata, pues, de comparar la v. dependiente entre cada uno de los dos grupos definidos sobre la v. independiente.

Las cuatro pruebas que incorpora este procedimiento se eligen en el mismo cuadro de diálogo tal como puede verse en la Figura 15.5 (a). El cuadro para definir los grupos de la variable independiente se muestra en la Figura 15.5 (b), y el cuadro de Opciones es el que se muestra en la Figura 15.5 (c), en el cual se puede seleccionar Estadísticos (número de casos, media, desviación típica, mínimo y máximo) y los tres cuarteles.

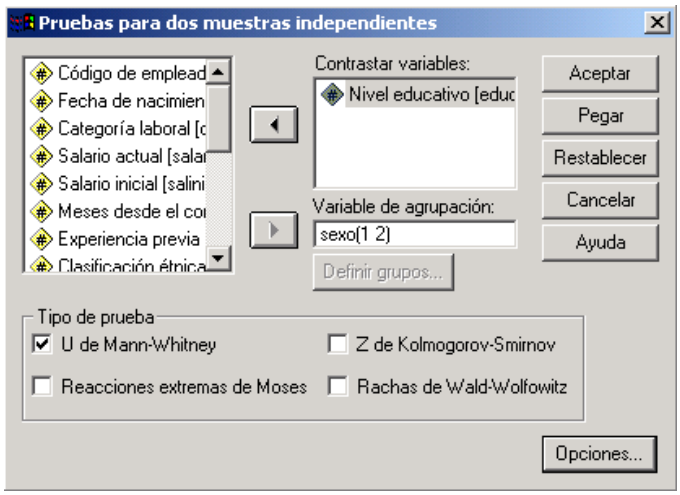


Figura 15.5 (a)

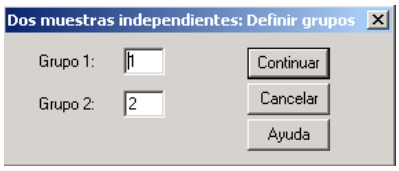


Figura 15.5 (b)

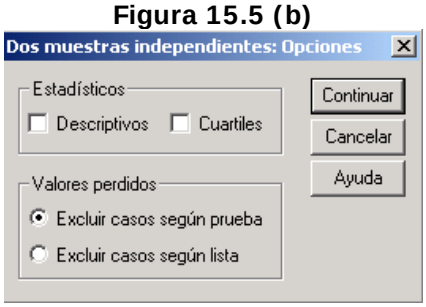


Figura 15.5 (c)

Figuras 15.5 (a). Cuadro de diálogo de Pruebas no paramétricas para dos muestras independientes; 15.5 (b) cuadro para definir los grupos de la v. independiente; 15.5 (c) cuadro de opciones del procedimiento.

15.3.1 Prueba U de Mann–Whitney

Esta prueba es el equivalente no paramétrico de la prueba T sobre diferencia de medias para muestras independientes, y se emplea cuando no se cumplen los supuestos de la prueba T (normalidad y homocedasticidad), o el nivel de medida de la variable es ordinal.

El contraste se realiza asignando rangos (R_i) ascendentes a las observaciones de los dos grupos una vez mezclados y ordenadas las observaciones. Hecho esto, se tendrá n_1 rangos (R_{i1}) correspondientes a las observaciones del grupo 1 y n_2 rangos (R_{i2}) correspondientes a las observaciones del grupo 2. Si sumamos los rangos para cada grupo, tendremos el valor S_1 = "suma de los rangos asignados a las observaciones del grupo 1" y el valor S_2 = "suma de los rangos asignados a las observaciones del grupo 2". Definimos ahora el estadístico U, que en cada grupo toma la siguiente forma

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - S_1 \qquad \text{y} \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - S_2$$

Si las poblaciones de las que se han extraído la dos muestras son iguales el valor de ambos estadísticos será aproximadamente igual. Si, por el contrario, las poblaciones de referencia no son iguales los valores de U_1 y U_2 serán también diferentes, y en estas condiciones de desigualdad se rechazaría la hipótesis de igualdad del promedio entre ambas poblaciones. Observe el lector que tal como está definido el estadístico U para cada grupo, un valor grande de U en uno de los grupos implica un valor pequeño de U en el otro grupo. La decisión sobre la probabilidad concreta asociada al estadístico U se puede basar en:

$$U = U_1 \quad \text{si } U_1 < n_1 n_2/2$$

$$U = U_2 \quad \text{si } U_1 > n_1 n_2/2$$

Con muestras menores de 30 (sumando n_1 y n_2) SPSS ofrece el nivel crítico bilateral exacto asociado al estadístico U, el cual se obtiene multiplicando por 2 la probabilidad de obtener valores menores o iguales que U. Con muestras mayores de 30, SPSS calcula una tipificación del estadístico U (incluyendo la corrección por empates) que se distribuye según el modelo de probabilidad normal N(0,1), y cuya expresión es:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{n(n-1)}\left(\frac{n^3 - n}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12}\right)}}$$

donde *k* se refiere al número de rangos distintos en los que se dan empates y t_i al número de puntuaciones empatadas en el rango *i*). El nivel crítico bilateral, se obtiene multiplicando por 2 la probabilidad de obtener valores menores o iguales que Z.

La tabla resultante de esta prueba se muestra en la Tabla 15.5. La variable de agrupamiento es el *sexo* y la variable dependiente es el nivel educativo (*educ*)

Tabla 15.5. Tablas de resultados de la prueba U de Mann–Whitney

Rangos

	Sexo	N	Rango promedio	Suma de rangos
Nivel educativo	Hombre	258	281,78	72699,50
	Mujer	216	184,61	39875,50
	Total	474		

Estadísticos de contraste [a]

	Nivel educativo
U de Mann-Whitney	16439,500
W de Wilcoxon	39875,500
Z	-8,031
Sig. asintót. (bilateral)	,000

a. Variable de agrupación: Sexo

15.3.2 Prueba de *reacciones extremas* de Moses

Con esta prueba se puede contrastar si existe diferencia en la dispersión de dos distribuciones, y por tanto permite el estudio de la variabilidad en sí misma, no como un supuesto de otro tipo de pruebas, tal como la comparación de medias.

Ya se ha visto un procedimiento de contrastar la hipótesis de homocedasticidad con la prueba de Levene (1960) que ofrece el procedimiento Explorar, pero este es un procedimiento paramétrico. La prueba de Moses (1952) puede usarse con variables ordinales. Las reacciones extremas se refieren a la situaciones en que los sujetos o bien están en la parte baja de la distribución o bien están en la parte alta.

Los grupos que se comparan con este procedimiento pueden considerarse como grupo de *control* (*c*) y como grupo *experimental* (*e*), que se suponen extraídas de

la misma población o de poblaciones idénticas. Para SPSS, el grupo de control siempre es el que el que tiene el menor valor. Para calcular el estadístico, se ordenan primero todas las observaciones ($n = n_c + n_e$) de forma ascendente como si se tratara de una única muestra, y se asigna rangos a cada observación, desde la más pequeña (rango 1) a la mayor (rango *n*). A continuación, se calcula la amplitud del grupo de control (A_c) restando los rangos correspondientes al valor más grande y más pequeño de este grupo, y sumando 1 a esa diferencia. El resultado se redondea al entero más próximo.

Para hacer más estable esta medida de dispersión, Moses sugiere utilizar la Amplitud recortada (A_r). Para ello, se fija un pequeño valor (*r*) y se calcula la amplitud tras eliminar del cómputo los *r* valores del grupo de control por arriba y por abajo (en total, 2*r*). Con los valores restantes se procede del mismo modo que para la amplitud sin recortes.

Procediendo de este modo, el valor de A_r tiene que estar en el intervalo n_c-2*r* y *n*-2*r*. Si en el grupo experimental se han producido reacciones extremas, la amplitud del grupo control tenderá a su valor mínimo, pues habrá pocos observaciones de este grupo entremezcladas con las del grupo control. La probabilidad de que el valor de A_r supere en alguna cantidad s, el valor n_c-2*r* viene dada por:

$$P\left(A_r \le n_c - 2r + s\right) = \frac{\sum_{i=0}^{s}\left[\binom{i + n_c - 2r - 2}{i}\binom{n_e + 2r + 1 + -i}{n_e - 1}\right]}{\binom{n}{n_c}}$$

Si la probabilidad es pequeña (inferior a 0,05) se puede rechazar la hipótesis de que ambas muestras proceden de poblaciones con la misma amplitud.

El resultado de esta prueba para contrastar la variabilidad de los años de educación (*educ*) se muestra en la Tabla 15.6

Tabla 15.6 Tablas de resultados de la prueba de *reacciones extremas* de Moses

Frecuencias

	Sexo	N
Nivel educativo	Hombre (Control)	258
	Mujer (Experimental)	216
	Total	474

Estadísticos de contraste [a,b]

		Nivel educativo
Amplitud observada del grupo control		448
	Sig. (unilateral)	,000
Amplitud recortada del grupo control		432
	Sig. (unilateral)	,749
Valores atípicos recortados de cada extremo		12

a. Prueba de Moses

b. Variable de agrupación: Sexo

15.3.3 Prueba de Kolmogorov–Smirnov para dos muestras

Esta prueba permite contrastar la hipótesis de que dos muestras proceden de la misma población, comparando las funciones de distribución de ambas muestras: $F_1(X_i)$ y $F_2(X_2)$. A diferencia de la prueba U de Mann–Whitney, que sólo compara promedios de rangos, esta prueba es sensible a cualquier diferencia entre las distribuciones, sea la tendencia central, la variabilidad, la simetría, etc.

Para realizar el contraste se empieza asignando rangos por separado para cada muestra para cada valor de X_i, por el procedimiento habitual para obtener las proporciones acumuladas. A continuación se obtienen las diferencias $D_i = F_1(X_i) - F_2(X_2)$, donde $F_1(X_i)$ se refiere a la función de distribución de la muestra de mayor tamaño. El estadístico de contraste es una tipificación de la diferencia más grande en valor absoluto:

$$Z_{K-S} = \text{máx}|D_i|\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

y se distribuye según el modelo de probabilidad normal N(0,1). Si la probabilidad de obtener una diferencia tan grande como la observada es inferior a 0,05 se rechaza la hipótesis de que ambas muestras proceden de la misma población.

El resultado de la prueba puede verse en la Tabla 15.7.

Tabla 15.7 Tablas de resultados de la *prueba de Kolmogoror-Smirnov* para dos muestras

Frecuencias

	Sexo	N
Nivel educativo	Hombre	258
	Mujer	216
	Total	474

Estadísticos de contraste [a]

		Nivel educativo
Diferencias más extremas	Absoluta	,402
	Positiva	,000
	Negativa	-,402
Z de Kolmogorov-Smirnov		4,359
Sig. asintót. (bilateral)		,000

a. Variable de agrupación: Sexo

15.3.4 Prueba de las rachas de Wald–Wolfowitz

Esta prueba es similar a la prueba de rachas para una muestra, y permite contrastar la hipótesis de que ambas muestras proceden de la misma población. Al igual que la prueba de K–S, esta prueba es sensibles a diferencias entre las muestras en cualquiera de las propiedades básicas de la distribución (centralidad, variabilidad, etc.).

Para obtener el número de rachas se ordenan de menor a mayor todas las observaciones (la suma de ambas muestras), y luego se obtienen el número de rachas (R) pertenecientes al mismo grupo. Si existen empates entre observaciones de muestras distintas, SPSS calcula tanto el número mínimo de rachas como el máximo.

Si las dos muestras proceden de la misma población, las observaciones de las dos muestras estarán entremezcladas y el número de rachas será alto. Si, por el contrario, las muestras provienen de poblaciones distintas una de ellas tendrá valores más alta que la otra y el número de rachas será bajo.

Para decidir sobre la significación estadística de si el número de rachas es lo suficientemente pequeño como para rechazar la hipótesis, SPSS lo hace de dos formas diferentes según sea la muestra total (suma de ambas muestras) menor o igual o mayor de 30. Si es mayor, ofrece una aproximación a la normal (igual que en la prueba de rachas para una muestra) pero con un nivel crítico correspondiente a un contraste unilateral.

Si es menor o igual de 30, SPSS ofrece un nivel crítico unilateral exacto. Si el número de rachas *R* es par, la ecuación que usa es la siguiente:

$$P(r \leq R) = \frac{2}{\binom{n}{n_1}} \sum_{r=2}^{R} \binom{n_1 - 1}{\frac{r}{2} - 1} \binom{n_2 - 1}{\frac{r}{2} - 1}$$

y si el número de rachas *R* es impar ($k = 2r - 1$),

$$P(r \leq R) = \frac{2}{\binom{n}{n_1}} \sum_{r=2}^{R} \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 2} + \binom{n_1 - 1}{k - 2} \binom{n_2 - 1}{k - 1}$$

Si la probabilidad obtenida es menor que 0,05 entonces se rechaza la hipótesis de que las muestras proceden de la misma población.

En la Tabla 15.8 se puede ver el resultado de la prueba.

Tabla 15.8 Tablas de resultados de la *prueba de rachas de Wald-Wolfowitz*

Frecuencias

	Sexo	N
Nivel educativo	Hombre	258
	Mujer	216
	Total	474

Estadísticos de contraste [b,c]

		Número de rachas	Z	Sig. asintót. (unilateral)
Nivel educativo	Mínimo posible	7[a]	-21,239	,000
	Máximo posible	288[a]	4,807	1,000

a. Hay 5 empates inter-grupos que implican 429 casos.

b. Prueba de Wald-Wolfowitz

c. Variable de agrupación: Sexo

La prueba ofrece el número mínimo y máximo de rachas, que depende del tratamiento que se de a los empates, el valor del estadístico Z para cada supuesto y el nivel crítico unilateral. Cuando el número de empates implican a muchos casos (como en el ejemplo) hay discrepancia entre el contraste para cada situación, y es conveniente emplear otra prueba para tomar una decisión sobre la hipótesis.

15.4 Pruebas para más de dos muestras independientes

Las pruebas de este procedimiento, son la alternativa no paramétrica al Anova de un factor, es decir, datos provenientes de diseños que contienen una variable categórica o factor con más de dos niveles y una variable dependiente que al menos esté medida a nivel ordinal. Incluye dos pruebas: la H de Kruskal–Wallis y la prueba de la Mediana.

En las Figuras 15.6 (a) a 15.6 (c) se muestran los cuadros y sub-cuadros de diálogo del procedimiento.

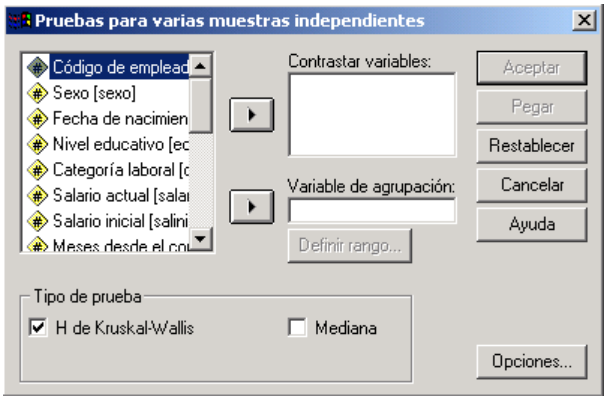


Figura 15.6 (a)

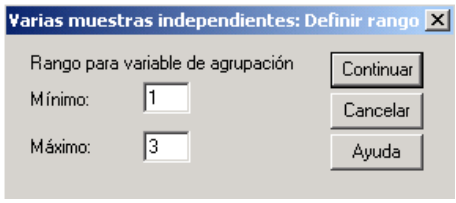


Figura 15.6 (b)

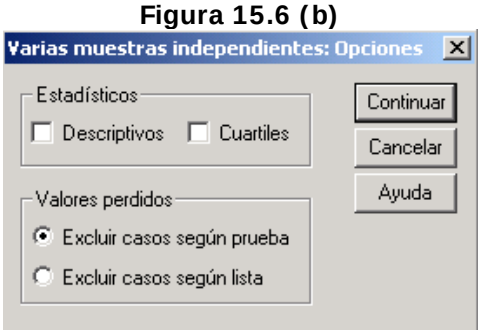


Figura 15.6 (c)

Figuras 15.6 (a). Cuadro de diálogo de *Pruebas no paramétricas para más de dos muestras independientes*; 15.6 (b). Cuadro para definir el *rango de la variable de agrupación*; 15.6 (c) Cuadro de *Opciones* del procedimiento

15.4.1 Prueba de Kruskal–Wallis

Esta prueba es similar al diseño del ANOVA de un factor completamente aleatorizado, en la cual se trata de contrastar si más de dos muestras aleatorias han sido extraídas de la misma población o de poblaciones distintas, sin necesidad de imponer restricciones sobre la distribución de la población, además de permitir trabajar con datos ordinales. No obstante, si se cumplen los supuestos para realizar un ANOVA, hay que señalar que el estadístico F presenta una mayor potencia que el estadístico H de esta prueba.

Para llevarla a cabo, en primer lugar hay que asignar rangos a las *n* observaciones que serán el resultado de sumar las n_1, n_2, ..., n_J. En caso de empates se asigna el promedio de los rangos empatados. Ahora llamamos R_{ij} a los rangos asignados a las *i* observaciones de la muestra *j*. Y llamamos R_j a la suma de los rangos asignados a las n_j observaciones de la muestra *j*. Es decir: $R_j = \sum_{i=1}^{n_j} R_{ij}$.

Si la hipótesis de que las muestras han sido extraídas de la misma población es verdadera, los R_j de cada muestra serán muy similares (salvo las diferencias atribuibles al azar, debidas al proceso de muestreo). El estadístico de esta prueba contrasta la hipótesis nula de que los rangos promedios de las J poblaciones son iguales, y lo hace a través de la suma de los rangos para cada grupo. Su expresión es:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{j} \frac{R_j^2}{n_j} - 3(n+1)$$

y se distribuye según el modelo de probabilidad chi-cuadrado con J-1 grados de libertad. Cuando hay rangos empatados, se emplea una corrección que hace el contraste más conservador. La corrección es:

$$H^* = \frac{H}{1 - \sum_{i=1}^{k} \frac{(t_i^3 - t_i)}{(n^3 - n)}}$$

siendo k el número de rangos distintos en los que existen empates, y t_i el número de valores empatados en cada rango.

Si se realizar esta prueba sobre la variable nivel educativo (educ), tomando la variable categoría laboral como variable de agrupamiento, el resultado que se muestra en el Visor es el de la Tabla 15.9.

Tabla 15.9 Tablas de resultados de la *prueba de Kruskal–Wallis* para más de dos muestras independientes

Rangos

	Categoría laboral	N	Rango promedio
Nivel educativo	Administrativo	363	206,43
	Seguridad	27	95,89
	Directivo	84	417,27
	Total	474	

Estadísticos de contraste [a,b]

	Nivel educativo
Chi-cuadrado	209,516
gl	2
Sig. asintót.	,000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Categoría laboral

15.4.2 Prueba de la mediana

Esta prueba es similar a la prueba chi-cuadrado ya estudiada en los procedimientos de análisis de datos categóricos. La diferencia estriba en que en esta prueba una de las variables (que será la variable dependiente) es al menos ordinal, y se dicotomiza utilizando la mediana como punto de corte.

Al tener una variable categórica (que definen *J* muestras), el objetivo de la pruebaes contrastar la hipótesis de que esas J muestras proceden de poblaciones con la misma mediana, para lo que se comienza ordenando todas las observaciones de la variable dependiente y calculando la mediana total, de acuerdo a:

$$Md = (X_{n/2} + X_{[n/2]+1}) \text{ si } n \text{ es par}$$

$$Md = (X_{[n+1]/2}) \text{ si } n \text{ es impar}$$

siendo X_n el valor más grande y X_1 el valor más pequeño. Seguidamente, se registra, en cada muestra, el número de casos con puntuación menor o igual que la mediana (grupo 1) y el número de casos con puntuación mayor que la mediana (grupo 2). Con el resultado ya se puede construir una tabla de contingencia 2 x J, siendo las dos filas las correspondientes a los dos grupos formados, y las J columnas las correspondientes a la J muestras independientes. Si las J muestras hubieran sido extraídas de la misma población, el número de casos por grupo en cada una de las J muestras sería aproximadamente el mismo, es decir, el 50 % del tamaño de cada muestra, y las frecuencias esperadas coincidirían con las frecuencias observadas.

El resultado de la prueba se muestra en la Tabla 15.10.

Tabla 15.10 Tablas de resultados de la Prueba de la *mediana* para más de dos muestras independientes

Frecuencias

		Categoría laboral		
		Administrativo	Seguridad	Directivo
Nivel educativo	> Mediana	147	1	83
	<= Mediana	216	26	1

Estadísticos de contraste [b]

	Nivel educativo
N	474
Mediana	12,00
Chi-cuadrado	116,082[a]
gl	2
Sig. asintót.	,000

a. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 13,2.

b. Variable de agrupación: Categoría laboral

15.5 Pruebas para dos muestras relacionadas

Estas pruebas permiten analizar datos provenientes de diseños con medidas repetidas, y son la alternativa no paramétrica de la prueba t para muestras relacionadas. De las tres pruebas que incluye el cuadro de diálogo, una de ellas, la de *McNemar* ya se ha visto en el capítulo correspondiente a análisis de datos categóricos, de modo que aquí sólo trataremos las otras dos.

El cuadro de diálogo y el correspondiente cuadro de Opciones se muestran en las Figuras 15.7 (a) y (b).

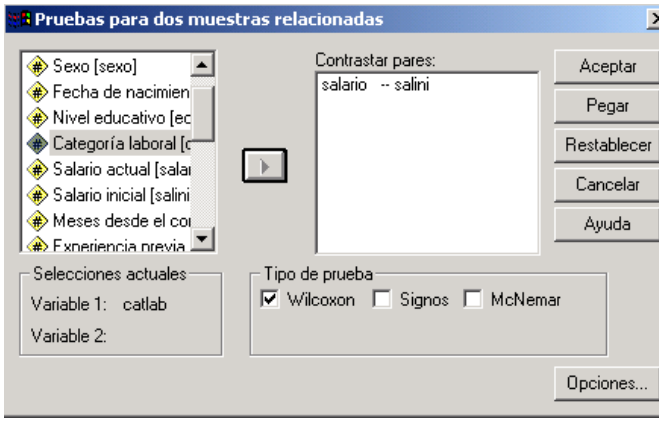


Figura 15.7 (a)

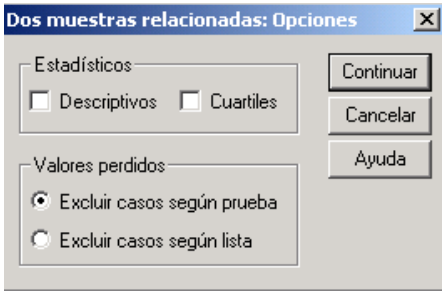


Figura 15.7 (b)

Figuras 15.7 (a). Cuadro de diálogo de *Pruebas no paramétricas para dos muestras relacionadas*; y 15.7 (b). *Opciones* de las pruebas

15.5.1 Prueba de Wilcoxon

Si se miden dos variables (X_i e Y_i) a un grupo de n sujetos, y se calculan las diferencias en valor absoluto entre las dos putnuaciones para cada par, se tiene otra variable D_i, que se expresa:

$$D_i = |X_i - Y_i| \qquad \text{para } i = 1, 2, \ldots, n$$

Si se consideran sólo las diferencias de las m diferencias no nulas, y se les asigna un rango ascendente, el 1 a la más pequeña y el m a la más grande (resolviendo los empates por el procedimiento del promedio), y posteriormente se suma, por un lado, los rangos positivos, R_i^+, es decir, los de los casos en que $X_i > Y_i$, y a la suma se le designa como S_+, y por otro, se suman los rangos negativos, R_i^-, es decir, los de los casos en que $X_i < Y_i$, y a la suma se le designa como S_-. Si, por último, suponemos que las puntuaciones X_i e Y_i, proceden de dos poblaciones con la misma mediana ($Md_X = Md_Y$), es esperable que:

$$P(X_i < Y_i) = P(X_i > Y_i)$$

lo cual, si se confirma, supone que deberemos encontrar tantos valores $X_i < Y_i$ como valores $X_i > Y_i$. Si, además, la distribución es simétrica, se puede esperar que:

$$S_+ = \sum R_i^+ = S_- = \sum R_i^-$$

Si esto es así, se confirmará la hipótesis nula de igualdad de que ambas muestras proceden de poblaciones con la misma mediana; en cambio, si hay discrepancia entre estos valores conducirá al rechazo de H_0.

Con tamaños muestrales pequeños es fácil obtener la distribución exacta del estadístico S (cuyo valor será el menor de S_+ o S_-), pero es más cómodo tipificar dicho valor, mediante:

$$Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{48}}}$$

que se distribuye según el modelo de probabilidad normal N(0,1), y en donde *k* es el número de rangos distintos en los que se producen empates, y t_i es el número de puntuaciones empatadas en el rango *i*. El nivel crítico que calcula SPSS es el resultado de multiplicar por 2 la probabilidad de obtener valores menores o iguales que Z.

El resultado de la prueba se muestra en la Tabla 15.11.

Tabla 15.11 Tablas de resultados de la Prueba de *Wilcoxon* para dos muestras relacionadas

Rangos

		N	Rango promedio	Suma de rangos
Salario inicial - Salario actual	Rangos negativos	474[a]	237,50	112575,00
	Rangos positivos	0[b]	,00	,00
	Empates	0[c]		
	Total	474		

a. Salario inicial < Salario actual

b. Salario inicial > Salario actual

c. Salario actual = Salario inicial

Estadísticos de contraste [b]

	Salario inicial - Salario actual
Z	-18,865[a]
Sig. asintót. (bilateral)	,000

a. Basado en los rangos positivos.

b. Prueba de los rangos con signo de Wilcoxon

15.5.2 Prueba de los signos

Esta prueba es más general que la prueba de Wilcoxon, y también permite contrastar la hipótesis de que dos muestras proceden de poblaciones con la misma mediana. Se diferencia, no obstante, en que en esta prueba sólo se aprovecha de los datos sus características nominales, aunque exija un nivel de medida al menos ordinal. Su fundamento es el siguiente: supongamos que se toman dos medidas (X_i e Y_i) a un grupo de *n* sujetos y que calculamos las diferencias

$$D_i = X_i - Y_i \quad \text{para } i = 1, 2, ..., n$$

entre las puntuaciones de cada para de observaciones. Si se descartan las diferencias nulas y se supone que las puntuaciones provienen de poblaciones con la misma mediana, entonces habrá el mismo número de diferencias positivas y negativas, es decir:

$$P(X_i < Y_i) = P(X_i > Y_i) = 0,5$$

Si esto se cumple, el número de diferencias positivas, n_+, y el número de diferencias negativas, n_-, se distribuirán con probabilidad la del modelo binomial con parámetros n y $n = 0,5$.

Si la muestra es pequeña ($n \leq 25$) el modelo binomial servirá para obtener la probabilidad asociada al valor k = mín (n_+, n_-). El nivel crítico será el que resulte de multiplicar por 2 la probabilidad de obtener valores iguales o menores que k. Si la muestra es mayor de 25, entonces se tipifica el valor de k (utilizando la corrección por continuidad), y se obtiene el valor:

$$Z = \frac{k + 0,5 - \frac{n}{2}}{0,5\sqrt{n}}$$

que se distribuye según el modelo de probabilidad normal N(0,1). El nivel crítico es el resultado de multiplicar por 2 la probabilidad de encontrar valores iguales o menores que Z.

El resultado de la prueba se muestra en la Tabla 15.12.

Tabla 15.12 Tablas de resultados de la Prueba de los *signos* para dos muestras relacionadas

Frecuencias

		N
Salario inicial - Salario actual	Diferencias negativas [a]	474
	Diferencias positivas [b]	0
	Empates [c]	0
	Total	474

a. Salario inicial < Salario actual

b. Salario inicial > Salario actual

c. Salario actual = Salario inicial

Estadísticos de contraste [a]

	Salario inicial - Salario actual
Z	-21,726
Sig. asintót. (bilateral)	,000

a. Prueba de los signos

15.6 Pruebas para más de dos muestras relacionadas

Las tres pruebas integradas en esta categoría, permiten el análisis de datos provenientes de diseños con medidas repetidas. Para ilustrar los procedimientos (excepto para la prueba de Cochran, que requiere datos dicotómicos o dicotomizados) utilizaremos los mismos datos empleados en el capítulo dedicado al ANOVA de un factor con medidas repetidas, datos que reproducimos a continuación:

	Nivel dificultad lectura			
Sujetos	1	2	3	4
1	14	12	7	6
2	15	10	9	9
3	16	8	11	9
4	13	11	8	9
5	16	12	7	12
6	16	10	8	11
7	14	13	12	10
8	12	8	11	7
9	11	8	8	10

El cuadro de diálogo de esta categoría de pruebas y el correspondientes a los estadísticas se muestran en las Figuras 15.8 (a) y 15.8 (b).

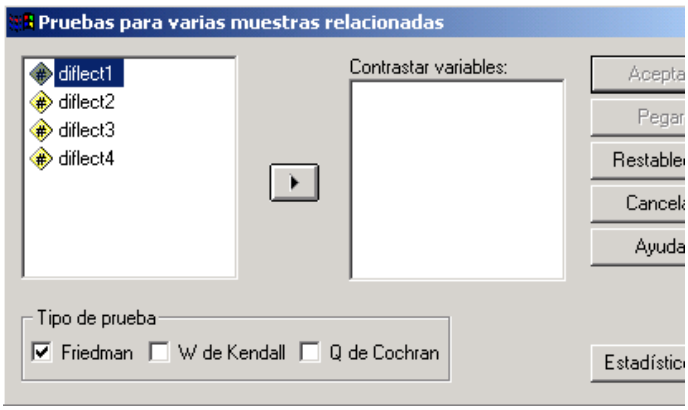


Figura 15.8 (a)

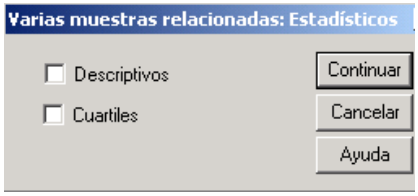


Figura 15.8 (b)

Figuras 15.8 (a). Cuadro de diálogo de Pruebas no paramétricas para más de dos muestras relacionadas; y 15.8 (b). Estadísticos de las pruebas

15.6.1 Pruebas de Friedman

Esta prueba es la alternativa no paramétrica a la del ANOVA de un factor con medidas repetidas, es decir, a aquellos diseños en donde a n sujetos (o n bloques) se les aplican J tratamientos y se les toma *J* medidas con el propósito de evaluar si los tratamientos son o no iguales.

Para llevar a cabo la prueba, hay que asignar rangos a las puntuaciones originales, pero de manera independiente para cada sujeto o bloque. Es decir, si hay *J* tratamientos, para cada sujeto se asignarán rangos de 1 a *J*. La suma de rangos para cada sujeto será siempre $J(J+1)/2$. Se designa como R_{ij} al rango asignado al sujeto (o bloque) *i* en el tratamiento *j*, y se designa como R_j la suma de los rangos asignados a todos los sujetos (o bloques) en el tratamiento j, es decir:

$$R_j = \sum_{i=1}^{n} R_{ij} \rightarrow \overline{R}_j = \frac{R_j}{n}$$

Si los promedios de las poblaciones son iguales todos los R_j serán similares. El estadístico de de Friedman contrasta si las J poblaciones (o tratamiento) son iguales, y su expresión es:

$$X_r^2 = \frac{12}{nJ(J+1)} \sum_j R_j^2 - 3n(J+1)$$

que se distribuye según el modelo de probabilidad *chi-cuadrado* con $J - 1$ grados de libertad.

El resultado de la prueba se muestra en la Tabla 15.13.

Tabla 15.13 Tablas de resultados de la Prueba de Friedman para más dos muestras relacionadas. Se incluyen las tablas con los estadísticos y los cuartiles

Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
DIFLECT1	9	14,11	1,833	11	16	12,50	14,00	16,00
DIFLECT2	9	10,22	1,922	8	13	8,00	10,00	12,00
DIFLECT3	9	9,00	1,871	7	12	7,50	8,00	11,00
DIFLECT4	9	9,22	1,856	6	12	8,00	9,00	10,50

Rangos

	Rango promedio
DIFLECT1	4,00
DIFLECT2	2,33
DIFLECT3	1,78
DIFLECT4	1,89

Estadísticos de contraste [a]

N	9
Chi-cuadrado	17,724
gl	3
Sig. asintót.	,001

a. Prueba de Friedman

15.6.2 Coeficiente de concordancia W de Kendall

Esta prueba sirve para estudiar la relación o acuerdo entre más de dos conjuntos de rangos. Cualquiera que sea la forma de obtener los rangos (clasificación de un grupo de jueces; clasificación según una determinada característica, etc.), se designa como R_{ij} al rango de sujeto *i* en la característica *j*, o al rango del sujeto *i* otorgado por el juez *j*, o..., y se designa como R_i a la suma de los rangos correspondientes al sujeto i.

$$R_i = \sum_{j=1}^{J} R_{ij}$$

La concordancia absoluta se producirá cuando todos los jueces clasifiquen a todos los sujetos de la misma manera, y la discordancia absoluta se producirá cuando los juicios de los jueces sean diferentes para cada sujeto. Cuando hay concordancia nula, todos los R_i serán iguales e igual a:

$$\frac{J(n+1)}{2}$$

Se ve, pues, que cuanto mayor sea la concordancia mayor será la variabilidad de los R_i, variabilidad que disminuirá a medida que el desacuerdo entre los jueces aumente. Entonces, se puede definir el estadístico S como

$$S = \sum_{i=1}^{n} \left(R_i - \frac{J(n+1)}{2} \right)^2$$

que es la variabilidad observada entre cada R_i y el total que cabría esperar si la concordancia fuera nula. El valor de S variará entre cero (desacuerdo máximo) y un valor máximo igual a:

$$S_{máx} = \frac{J^2 n(n^2 - 1)}{12}$$

que se obtendrá cuando el acuerdo sea total. Un coeficiente, que se denomina $\hat{W}$, que oscile entre 0 y 1, y que indique un desacuerdo máximo (0) o un acuerdo total (1) es el que se consigue dividiendo el valor de S entre el valor máximo de S:

$$\hat{W} = \frac{12\sum_i R_i^2}{J^2 n(n^2 - 1)} - \frac{3(n+1)}{n-1}$$

$\hat{W}$ se puede transformar en el estadístico X_r^2 de Friedman mediante

$$X_r^2 = J(n-1)\hat{W}$$

por lo que la hipótesis nula que se contrasta en Friedman de que los tratamientos son iguales es exactamente la misma que la de la ausencia de concordancia.

Cuando dentro de un mismo conjunto de rangos se produce un empate, SPSS realiza una corrección del estadístico que lo hace más conservador y su expresión es:

$$\hat{W} = \frac{12\sum_i R_i^2 - 3J^2 n(n+1)^2}{J^2 n(n^2-1) - J\sum_{i=1}^{k}(t_i^3 - t_i)}$$

donde k es el número de rangos distintos en los que se produce empate y t_i es el número de puntuaciones empatadas dentro de cada rango.

El resultado de la prueba se muestra en la Tabla 15.14.

Tabla 15.14 Tablas de resultados de la Prueba W de Kendall para más dos muestras relacionadas.

Rangos

	Rango promedio
DIFLECT1	4,00
DIFLECT2	2,33
DIFLECT3	1,78
DIFLECT4	1,89

Estadísticos de contraste

N	9
W de Kendall [a]	,656
Chi-cuadrado	17,724
gl	3
Sig. asintót.	,001

a. Coeficiente de concordancia de Kendall

15.6.3 Prueba de Cochran

Cuando a n sujetos se les toman J medidas de una variable dicotómica, se está en la misma situación que en un diseño ANOVA con medida repetidas o bloques con un sujeto por nivel o bloque, pero con la variable dependiente de naturaleza dicotómica.

Si llamamos aciertos al valor 1 y error al valor 0, las proporciones marginales P_{+j} de cada muestra o tratamiento será igual a $\frac{T_{+j}}{n}$, siendo T_{+j} la suma de aciertos en cada muestra. Si las J muestras proceden de poblaciones idénticas, las proporciones marginales P_{+j} serán aproximadamente iguales. Tomando este razonamiento como punto de partida, Cochran diseño un estadístico (*Q*) que permite contrastar la hipótesis de igualdad entre las J proporciones poblacionales (H_0: $\pi_{+1} = \pi_{+2} = ... = \pi_{+J}$), cuya expresión es:

$$Q = \frac{J(J-1)\sum T_{+j}^2 - (J-1)T^2}{JT - \sum T_{i+}^2}$$

que se distribuye según el modelo de probabilidad *chi-cuadrado* con *J*–1 grados de libertad. El nivel crítico que ofrece SPSS, es la probabilidad de obtener valores de *Q* iguales o mayores que el encontrado

Apéndice 1. Lectura de archivos de formato diferente a SPSS

A1.1 Introducción

Además de los archivos con formato SPSS (archivos con extensión .sav), es posible leer una variedad de formato de archivos de datos que cubre casi todas las necesidades del analista. Los formatos de archivos de datos que pueden leerse con SPSS son los siguientes:

SPSS. Abre archivos de datos guardados con formato SPSS, incluyendo SPSS para Windows, Macintosh, UNIX y el producto SPSS/PC+ para DOS.

SPSS/ PC+ . Abre archivos de datos de SPSS/PC+.

SYSTAT. Abre archivos de datos de SYSTAT.

SPSS portátil. Abre archivos de datos guardados con formato SPSS portátil. El almacenamiento de archivos en este formato lleva mucho más tiempo que guardarlos en formato SPSS.

Excel. Abre archivos de Excel.

Lotus 1-2-3. Abre archivos de datos guardados en formato 1-2-3 en las versiones 3.0, 2.0 ó 1A de Lotus.

SYLK. Abre archivos de datos guardados en formato SYLK (vínculo simbólico), un formato utilizado por algunas aplicaciones de hoja de cálculo.

dBASE. Abre archivos con formato dBASE para dBASE IV, dBASE III o III PLUS, o dBASE II. Cada caso es un registro. Las etiquetas de valor y de variable y las especificaciones de valores perdidos se pierden si se guarda un archivo en este formato.

Respecto a estos archivos sólo comentaremos algunos aspecto de los archivos de **Excel**, de los archivos de dBase, y de los archivos de texto en formato ASCII

A1.2 Lectura de archivos de Excel

En relación con los archivos de Excel. Cuando se desea abrir archivos de este tipo, hay que seleccionar dicho tipo en el menú de persiana "Tipo de archivo" del cuadro de diálogo "Abrir archivo". Elegido y abierto el archivo, se muestra el cuadro de diálogo de la Figura A1.1.

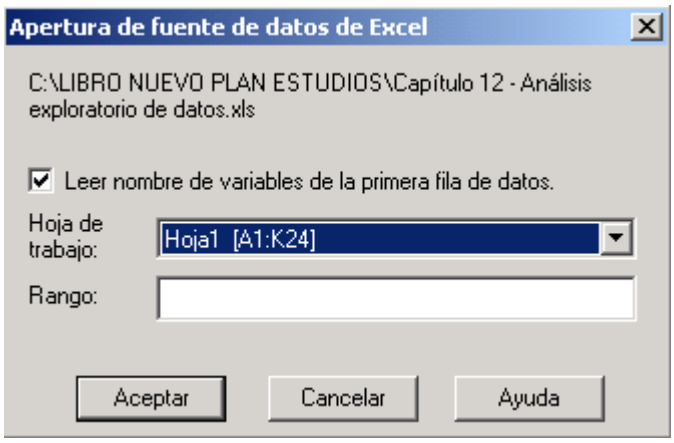


Figura A1.1 Cuadro de diálogo al abrir archivos de Excel

Para leer estos archivos correctamente, es preciso señalar la opción Leer nombres de variables, para que los nombres que aparecen en la primera fila de la hoja de cálculo, que es donde suelen situarse los nombres de las variables, los lea como nombres de variables para SPSS. Además, hay que señalar la hoja de cálculo que se quiere leer de todas las que componen un Libro de Excel. Para leer todas las hojas que componen un libro será necesario crear tantos archivos como hojas de cálculo tenga. Por último, si se desea sólo leer un rango de datos de la hoja de cálculo elegida, se especifica en el cuadro correspondiente; si no se señala un rango, se leen todos los datos que hay en la hoja de cálculo.

A1.3 Lectura de archivos de dBase

Se pueden leer archivos de base de datos de dBase II, III, III+ y IV. Al abrir un archivo de estos tipos los registros se convierten en casos y los campos en variables. Dado que en dBase los campos pueden nombrarse con hasta diez caracteres, SPSS los trunca a los ocho primeros. Si una variable después de truncado su nombre coincide con el nombre de alguna de las variables previas del archivo, SPSS le asigna el nombre por defecto VAR0001. Los registros marcados en dBase para ser borrados, pero que aún no lo han sido se leen como casos válidos. Para estos casos SPSS, genera una variable de cadena nueva con nombre D_R y asigna asteriscos a los casos que hubieran estado marcados para borrar. Si no hay ningún registro marcado, esta nueva variable se puede borrar.

A1.4 Lectura de archivos de texto

La opción Leer datos de texto del menú Archivo permite leer archivos de texto con fomato ASCII estándar, que son archivos que contienen caracteres, espacios en blanco y retornos de carro. Cuando se elige esta opción se accede a un asistente de 6 pasos que ayuda a leer los datos para transformarlos en un archivo de datos de SPSS. Para ilustrar el proceso vamos a utilizar un archivo de texto cuyos primeros casos se muestran en el cuadro siguiente:

```
1 1 1 6,7
1 1 2 4,3
1 1 3 5,7
1 2 1 7,1
. . . . .
```

Este archivo en formato ASCII estándar, que hemos nombrado como Pan.dat, contiene cuatro variables: **grasa**, **reactivo**, **harina**, y **volesp** (volumen específico). La variable grasa ocupa la columna 1, reactivo la columna 3, harina la columna 5 y volesp las columnas 7 a 9. Hay que señalar que cuando los datos no están delimitados por un carácter concreto (p.e., coma) hay que saber qué columna ocupa cada variable. Cuando abrimos el archivo, se abre el primer paso del asistente como en la Figura A1.2.

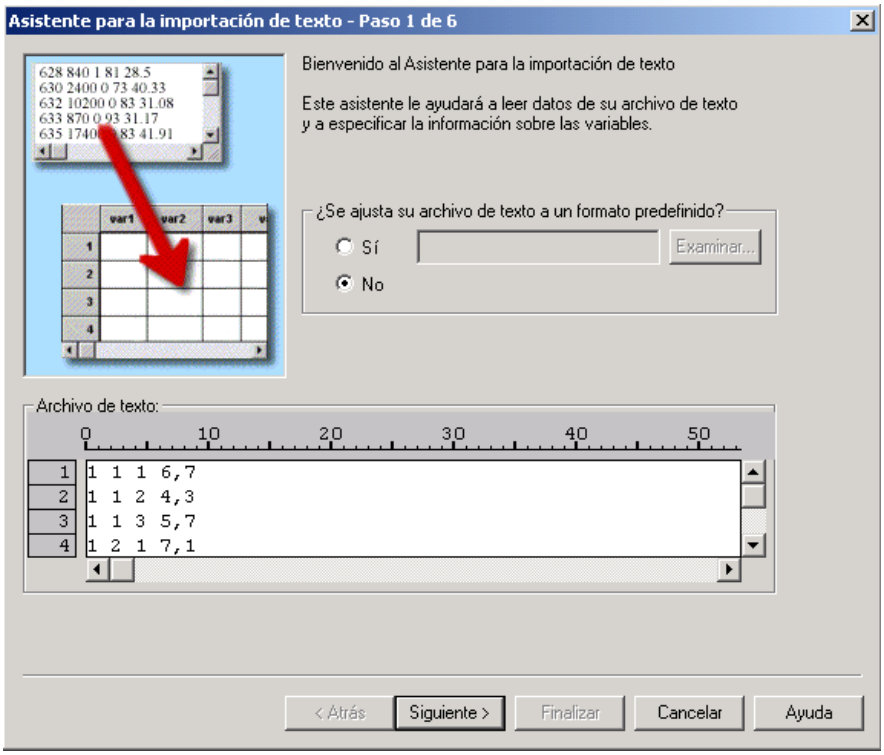


Figura A1.2 Cuadro de diálogo Asistente para importación de texto Paso 1 de 6

En este primer paso, hay que especificar si nuestro archivo se ajusta a un formato predefinido (que hayamos guardado previamente con este asistente) o no se ajusta. Elegida la opción, se pulsa Siguiente>, y se accede al **Paso 2** (Figura A1.3)

Hay que especificar cómo están organizadas las variables:

Delimitadas: separadas por espacios, comas tabulaciones u otros caracteres, o son de ancho fijo. En nuestro caso, podemos elegir cualquiera de las dos opciones, y el resultado será el mismo. Sin embargo cuando las variables estén en el archivo de texto sin espacios en blanco intermedios, sólo se pueden leer correctamente señalando la opción Ancho fijo. También hay que especificar si los nombres de las variables están en la parte superior del archivo, lo que es el caso en este ejemplo.

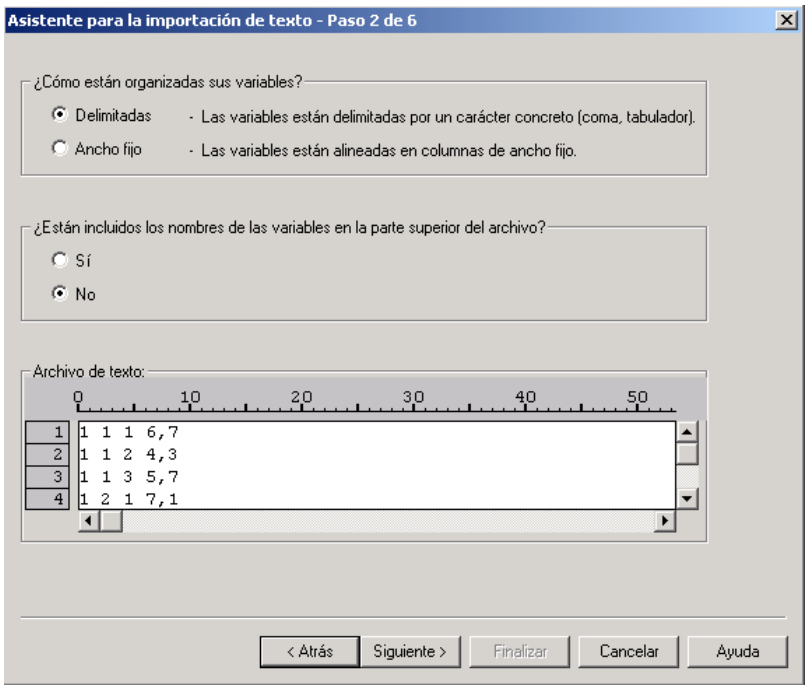


Figura A1.3 Cuadro de diálogo Asistente para importación de texto Paso 2 de 6

Se pulsa el botón Siguiente> y se accede al **Paso 3** (Figura A1.4)

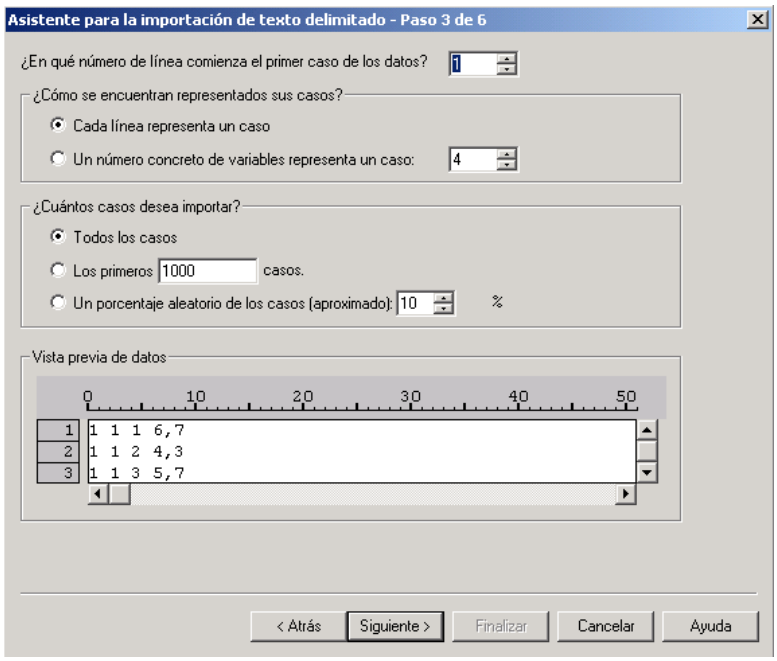


Figura A1.4 Cuadro de diálogo Asistente para importación de texto Paso 3 de 6

En este paso hay que completar la información de todos los campos. En primer lugar especificar el número de línea del archivo donde comienza el primer caso; después si cada caso ocupa una sola línea u ocupa más de una línea. Esto suele suceder cuando el numero de variables es muy grande y no caben todos las datos de un caso en una sola línea. A continuación hay que especificar si se leen todos los casos (opción por defecto) o sólo se leen un número de ellos o un porcentaje.

Hechas estas especificaciones se pulsa Siguiente> y se accede al **Paso 4** (Figura A1.5)

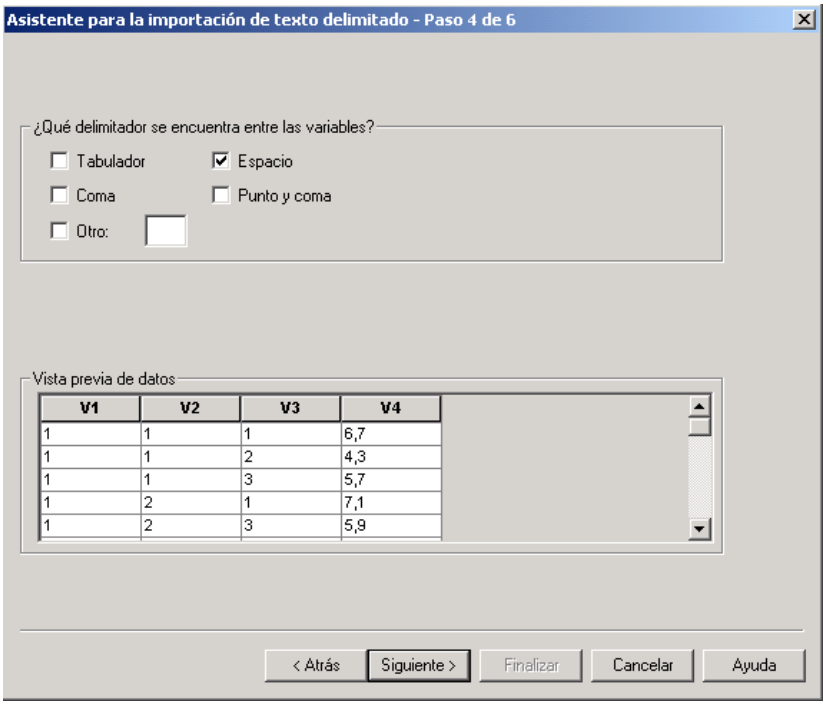


Figura A1.5 Cuadro de diálogo Asistente para importación de texto Paso 4 de 6

Como nuestro archivo sólo contiene espacio entre las variables, con solo especificar esta opción, el asistente organiza el archivo con esta información. Si en vez de espacio se hubiera señalado Tabulador, Coma o Punto y coma el resultado hubiera sido diferente al correcto (dejamos al lector que explore el resto de las opciones de este paso). A continuación se pulsa Siguiente> y se accede al **Paso 5** (Figura A1.6)

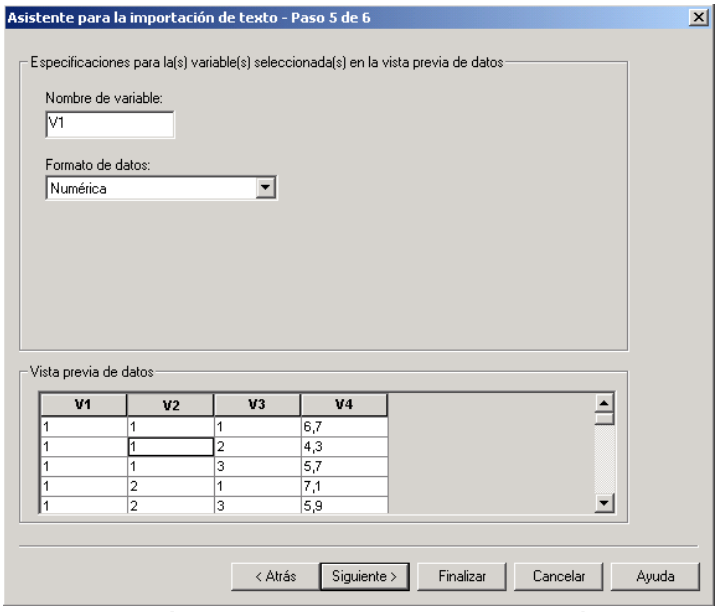


Figura A1.6 Cuadro de diálogo Asistente para importación de texto Paso 5 de 6

En este paso, se puede cambiar el nombre de las variables que el asistente asigna por defecto y el tipo de variable; incluso se puede elegir no importar la

variable. Cuando el dato es un número, el asistente asigna por defecto un formato de datos Numérico, y si encuentra letras asigna un formato de datos de Cadena. Para cambiar tanto el nombre como el formato, es preciso marcar la cabecera de la variable que se quiere cambiar y entonces se activa los cuadros Nombre de variable y formato de datos. El cambio de nombres y formato puede realizarse en este paso, pero las posibilidades que hay respecto a estos aspectos de la definición de variables en el propio editor de datos de SPSS son mayores que las que ofrece este asistente. En cualquier caso, se nombre o no en este paso las variables, siempre se termina pulsando Siguiente> y se accede al **Paso 6** (Figura A1.7)

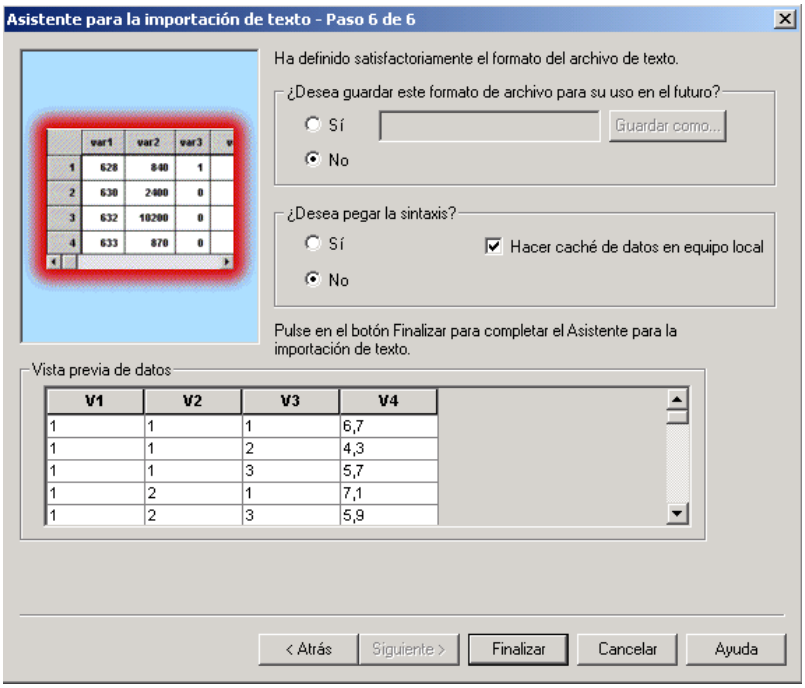


Figura A1.7 Cuadro de diálogo Asistente para importación de texto Paso 6 de 6

Después del Paso 5 ya se ha culminado todo lo relacionado con la definición del archivo de texto. No obstante, el asistente permite dos últimos controles que pueden resultar útiles para posteriores sesiones de lectura de datos de texto. Con la primera opción "¿Desea guardar este formato de archivo para su uso futuro?" se puede crear un archivo con todas las especificaciones que se han realizado para leer el archivo, de modo que puedan ser utilizadas en sesiones posteriores con archivos de igual formato, sin tener que repetir todos los pasos, con solo invocar el archivo en el primer paso del asistente (ver opción en Figura A1.2).

La otra opción es la posibilidad de Pegar la sintaxis en una ventana de sintaxis, desde la cual se hará la lectura del archivo de texto. Si se pulsa No en esta opción, cuando se pulsa el botón Finalizar, SPSS comienza la lectura del archivo y construye el archivo de datos en el Editor de datos. Si se pulsa Si, al pulsar Finalizar SPSS no leerá el archivo, y será preciso hacerlo desde la ventana de sintaxis.

A1.5 Cuando los archivos no tienen espacios en blanco

Si los archivos de texto no tienen espacios en blanco, en el paso 2 hay que especificar que los datos son de Ancho fijo y, en el paso 4, hay una diferencia

respecto a lo visto cuando los datos están *Delimitados por caracteres*. En la Figura A1.8 se ve este paso 4 y cómo hay que proceder.

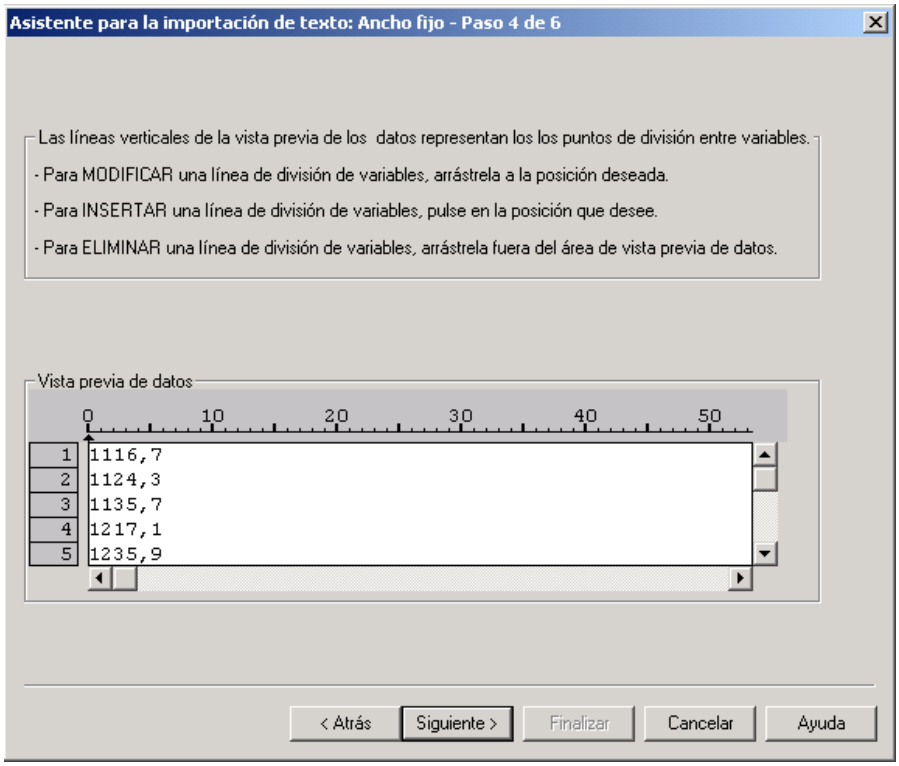


Figura A1.8 Cuadro de diálogo *Asistente para importación de texto Paso 4 de 6* cuando los datos son de Ancho fijo.

En este paso hay que especificar, mediante inserción de líneas, cuales son las variables. En nuestro caso, al ser cuatro las variables, emplearíamos tres líneas: una entre las columnas 1 y 2; otra entre las columnas 2 y 3; otra entre las columnas 3 y 4; y por último otra entre las columnas 4 y 5. Para insertar las líneas, se sitúa el puntero del ratón entre dos columnas y se hace clic. En caso de que insertemos una línea no deseada, se vuelve a pulsar sobre ella y se arrastra, con el ratón pulsado, fuera del cuadro donde aparecen los valores de las variables. Después de insertadas las líneas el aspecto del cuadro sería el de la Figura A1.9.

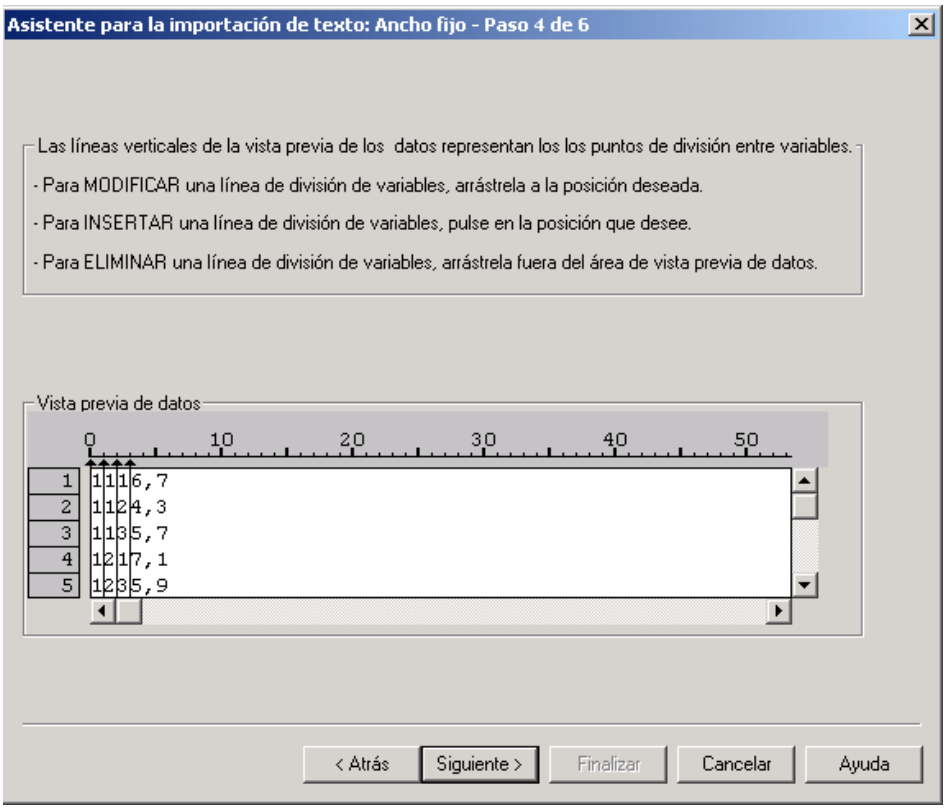


Figura A1.8 *Paso 4 de 6* del asistentecuando los datos son de Ancho fijo y ya se han insertado las líneas de división para identificar las variables.

Apéndice 2 Módulo de Tablas

A2.1 Introducción

La opción Tablas de SPSS incluye tres procedimientos que permite generar casi todo tipo de tablas con información resumen. Los tres tipos son: *Tablas básicas*, *Tablas generales* y *Tablas de frecuencia*.

En este apéndice vamos a describir de manera resumida cómo generar cada uno de los tipos, sin entrar en el detalle de cómo manipular una tabla cuando ya ha sido generada, porque ya se ha tratado en este manual.

El archivo de datos que vamos a emplear para ilustrar la generación de tablas es el denominado "*Encuesta general USA 1991.sav*", presente en el directorio en el que se ha instalado la aplicación. Esta encuesta fue dirigida por el Centro Nacional de Investigación de Opiniones y se realiza en ese país desde 1972.

A2.2 Estructura general de las tablas

Como ya conoce el lector las tablas puede ser de una, dos o tres dimensiones. Cada dimensión está definida por una variable o un conjunto de variables. Las variables que se muestran en la parte izquierda de la tabla se denominan variables de fila, las que aparecen en la parte superior se denominan variables de columna, y las que aparecen en tablas apiladas se denominan variables de capa, y son las que definen las tercera dimensión de la tabla. En la Figura A2.1 se observa las tres dimensiones.

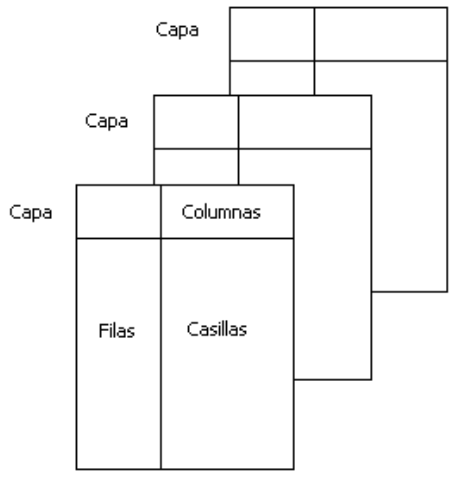


Figura A 2.1 Dimensiones físicas de una tabla

Si se sitúan varias variables en una misma dimensión, se pueden organizar de dos maneras diferentes: apiladas o anidadas. Si se apilan las categorías de cada variable aparecen separadas, primero las de una y después las de otra; si se anidan aparecen todas las categorías de la variable anidada dentro de cada categoría de la

variable que está por encima. En las Figuras A2.2(a) y A2.2 (b), se observa claramente las diferencias.

		Nivel de felicidad			¿Su vida es excitante o aburrida?		
		Muy feliz	Bastante feliz	No demasiado feliz	Excitante	Rutinaria	Aburrida
Sexo del encuestado	Hombre	206	374	53	213	200	12
	Mujer	261	498	112	221	305	29
Raza del encuestado	Blanca	409	730	117	371	413	34
	Negra	46	116	39	51	69	6
	Otra	12	26	9	12	23	1

Figura A2.2(a). Tabla con dos variables de fila (Sexo y Raza) y dos de columna (nivel de felicidad y percepción de su vida), apiladas en cada dimensión

				Nivel de felicidad								
				Muy feliz			Bastante feliz			No demasiado feliz		
				¿Su vida es excitante o aburrida?			¿Su vida es excitante o aburrida?			¿Su vida es excitante o aburrida?		
				Excitante	Rutinaria	Aburrida	Excitante	Rutinaria	Aburrida	Excitante	Rutinaria	Aburrida
Sexo del encuestado	Hombre	Raza del encuestado	Blanca	84	33		92	125	4	5	14	6
			Negra	8	3	1	10	14		7	2	
			Otra	3	1		3	5		1	1	
	Mujer	Raza del encuestado	Blanca	88	51	1	96	154	8	6	30	14
			Negra	10	8		14	31		2	10	5
			Otra	2	2		3	9			4	1

Figura A2.2(b). Tabla con las misma variables que en (a), pero anidadas en vez de apiladas.

Una vez generada la tabla, siempre se puede modificar su aspecto mediante el editor de tablas, al cual se accede pulsando dos veces consecutivas sobre la tabla en el *Visor* de SPSS. Cuando se está en el editor, se pueden activar los denominados paneles de pivotado y, si se desea, se procede al intercambio de las dimensiones de la tabla. También se puede modificar el aspecto, rotar las etiquetas para que una tabla se estreche y pueda entrar en el ancho de una página, o se puede decidir en que lugar cortar una tabla larga, o que elementos de la tabla deben permanecer juntos dentro de la página impresa.

A 2.3 Selección del tipo de tabla apropiado

Al disponer de tres tipos de tablas, es preciso, previamente, decidir qué tipo de tabla es la adecuada a las necesidades, aunque todas comparten elementos comunes, tales como la incorporación de títulos, notas a pie de página, etc.

En general, en la mayoría de las ocasiones las **Tablas básicas** son suficientes para nuestros propósitos. Este tipo de tablas contienen la mayor parte de los elementos necesarios para confeccionar una tabla que resuma adecuadamente la información. Además, su cuadro de diálogo es muy sencillo y las herramientas de que dispone se aplican de la misma manera a todas las variables que se seleccionen. Permite el apilamiento y anidamiento de variables, la incorporación de variables resumen (variables métricas de las que se quiere extraer información agregada: media, mediana, desviación típica, etc.), y la inserción de totales por fila y/o columna.

Cuando se quiere generar una tabla donde el apilado o anidado se aplica de manera diferente a según qué variable, o se quiere manejar variables de respuesta múltiple, hay que utilizar las **Tablas generales**. Como ejemplo, en la Figura A2.3 se muestra una tabla con las variables de fila apiladas y las variables de columna anidadas, lo cual sólo se consigue con este procedimiento.

		Región de los Estados Unidos								
		Nor-Este			Sur-Este			Oeste		
		Nivel de felicidad			Nivel de felicidad			Nivel de felicidad		
		Muy feliz	Bastante feliz	No demasiado feliz	Muy feliz	Bastante feliz	No demasiado feliz	Muy feliz	Bastante feliz	No demasiado feliz
Sexo del encuestado	Hombre	78	176	26	70	94	12	58	104	15
	Mujer	107	236	50	79	121	35	75	141	27
Raza del encuestado	Blanca	162	361	55	123	150	31	124	219	31
	Negra	18	44	19	23	56	14	5	16	6
	Otra	5	7	2	3	9	2	4	10	5

Figura A2.3 Tabla generada con el procedimiento Tablas generales con variables apiladas y anidadas

Por último, el procedimiento **Tablas de frecuencias** se utiliza para propósitos especiales, tales como mostrar frecuencias de variables que están categorizadas de la misma manera. En la Figura A2.4 se muestra un ejemplo.

	Obedecer es	Ayudar a otros es	Trabajar duro es
	Recuento	Recuento	Recuento
Lo más importante	195	126	147
Lo 2º más importante	123	316	355
Lo 3º más importante	142	332	321
Lo 4º más importante	343	175	144
Poco importante	179	33	15

Figura A2.4 Tabla de frecuencias

Como criterio general, en la tabla siempre aparece la etiqueta de las variables, y las etiquetas de los valores. Si una variable no tiene etiqueta se imprime el nombre de la variable, y si los valores no tienen etiqueta se imprimen los valores. Otro aspecto que comparten los diferentes procedimientos de tablas son el **Título** y los textos a **Pie** de tabla. El cuadro de diálogo es el que se muestra en la Figura A2.5.

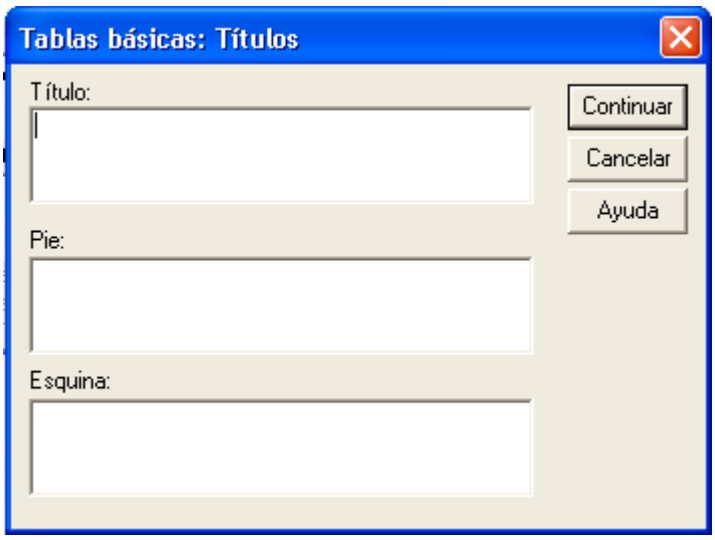


Figura A2.5 Cuadro de diálogo de Títulos, cuyo aspecto comparten las tablas básicas, las generales y las de frecuencia.

A2.4 Tablas básicas

El cuadro de diálogo de este procedimiento es el que se muestra en la Figura A2.6. En las tres listas englobadas bajo el término **Subgrupos** (Hacia abajo; A través; Tablas distintas), se incorporan las variables categóricas. Si las variables se incorporan a la lista **Hacia abajo**, las categoría de la variables aparecerán como filas; si se incorporan a la lista **A través**, las categorías aparecerán como columnas, y si se incorporan a la lista **Tablas distintas**, cada categoría se mostrará en tablas diferentes (las denominadas capas).

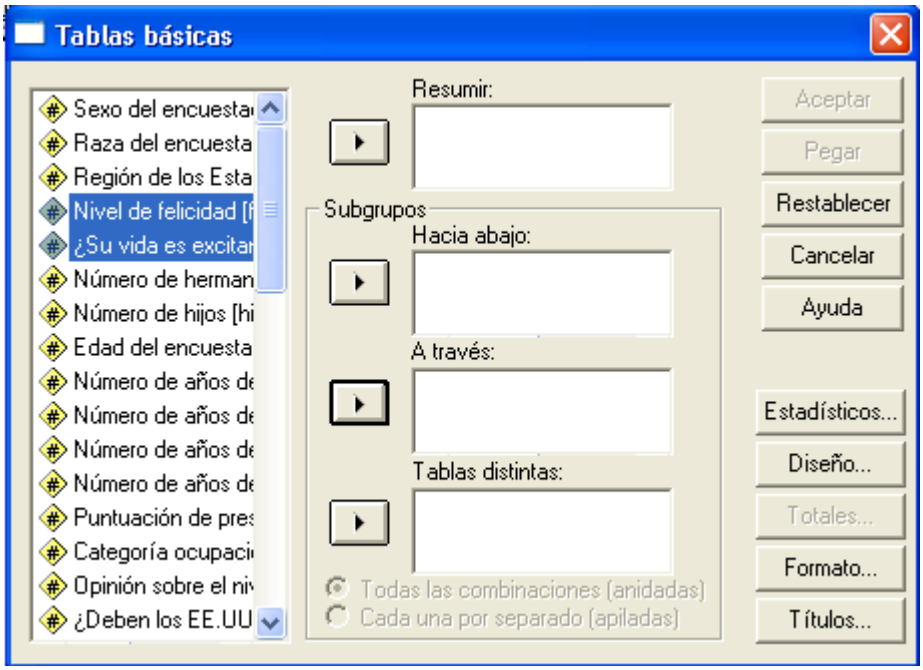


Figura A 2.6 Cuadro de diálogo de Tablas básicas

En la lista denominada **Resumir**, se incorporan las variables métricas sobre las que se quiere extraer información resumen. Todos los estadísticos incorporados en el procedimiento se pueden ver pulsando el botón **Estadísticos...**

Como ilustración del resultado mostramos, en la Figura A2.7 diferentes tablas con incorporación paulatina de variables y de la información que contiene la tabla.

Nivel de felicidad	Muy feliz	467
	Bastante feliz	872
	No demasiado feliz	165

Tabla con una variable en fila

		Región de los Estados Unidos		
		Nor-Este	Sur-Este	Oeste
Nivel de felicidad	Muy feliz	185	149	133
	Bastante feliz	412	215	245
	No demasiado feliz	76	47	42

Tabla con una variable en fila y otra en columna

Apéndice 2. El módulo de Tablas

		Región de los Estados Unidos		
		Nor-Este	Sur-Este	Oeste
Nivel de felicidad	Muy feliz	185	149	133
	Bastante feliz	412	215	245
	No demasiado feliz	76	47	42
Raza del encuestado	Blanca	582	307	375
	Negra	82	94	28
	Otra	15	14	20

Tabla con dos variables apiladas en fila y una variable en columna

				Región de los Estados Unidos		
				Nor-Este	Sur-Este	Oeste
Nivel de felicidad	Muy feliz	Raza del encuestado	Blanca	162	123	124
			Negra	18	23	5
			Otra	5	3	4
	Bastante feliz	Raza del encuestado	Blanca	361	150	219
			Negra	44	56	16
			Otra	7	9	10
	No demasiado feliz	Raza del encuestado	Blanca	55	31	31
			Negra	19	14	6
			Otra	2	2	5

Tabla con dos variables anidadas en fila y una en columna

				Región de los Estados Unidos			Total de tabla
				Nor-Este	Sur-Este	Oeste	
Raza del encuestado	Blanca	Sexo del encuestado	Hombre	248	138	159	545
			Mujer	334	169	216	719
	Negra	Sexo del encuestado	Hombre	28	33	10	71
			Mujer	54	61	18	133
	Otra	Sexo del encuestado	Hombre	5	6	9	20
			Mujer	10	8	11	29
Total de tabla				679	415	423	1517

Tabla con dos variables anidadas en fila y una en columna, incorporando el recuento Total de tabla

				Región de los Estados Unidos									Total de tabla
				Nor-Este			Sur-Este			Oeste			
				Sexo del encuestado		Total de grupo	Sexo del encuestado		Total de grupo	Sexo del encuestado		Total de grupo	
				Hombre	Mujer		Hombre	Mujer		Hombre	Mujer		
Nivel de felicidad	Muy feliz	Raza del encuestado	Blanca	69	93	162	57	66	123	54	70	124	409
			Negra	6	12	18	10	13	23	3	2	5	46
			Otra	3	2	5	3		3	1	3	4	12
		Total de grupo		78	107	185	70	79	149	58	75	133	467
	Bastante feliz	Raza del encuestado	Blanca	160	201	361	70	80	150	96	123	219	730
			Negra	14	30	44	21	35	56	3	13	16	116
			Otra	2	5	7	3	6	9	5	5	10	26
		Total de grupo		176	236	412	94	121	215	104	141	245	872
	No demasiado feliz	Raza del encuestado	Blanca	19	36	55	10	21	31	8	23	31	117
			Negra	7	12	19	2	12	14	4	2	6	39
			Otra		2	2		2	2	3	2	5	9
		Total de grupo		26	50	76	12	35	47	15	27	42	165
Total de tabla				281	398	679	177	238	415	178	245	423	1517

Tabla con dos variables anidadas en fila y otras dos anidadas en columna, con el total de grupo (total para cada variable) de grupo y el total de la tabla

			Región de los Estados Unidos		
			Nor-Este	Sur-Este	Oeste
			Media	Media	Media
Sexo del encuestado	Hombre	Número de años de escolarización	13	13	13
	Mujer	Número de años de escolarización	13	12	13

Tabla con una variable de fila y otra de columna y en las casillas información sobre el promedio de años de educación para cada cruce de categorías

Figura A2.7. Diferentes tipos de tablas generadas con el procedimiento de Tablas básicas

Cuando no se incorporan variables métricas para ser resumidas, los estadísticos que pueden integrar los casillas se refieren a recuentos, porcentajes sobre filas, porcentajes sobre columnas, porcentajes sobre tabla, etc. También, como se ha señalado, se pueden incorporar totales respecto de cada variable en cada dimensión, y totales respecto de la tabla. En las Figuras A2.8(a) y A2.8(b) se muestran los cuadros de diálogo para los estadísticos y los totales.

Figura A2.8(a) Cuadro de diálogo de Estadísticos de Tablas básicas

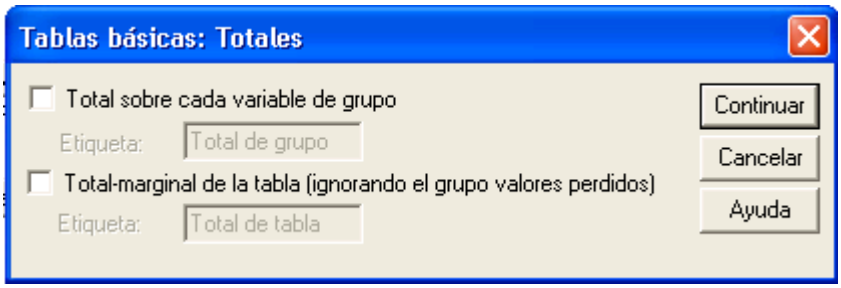
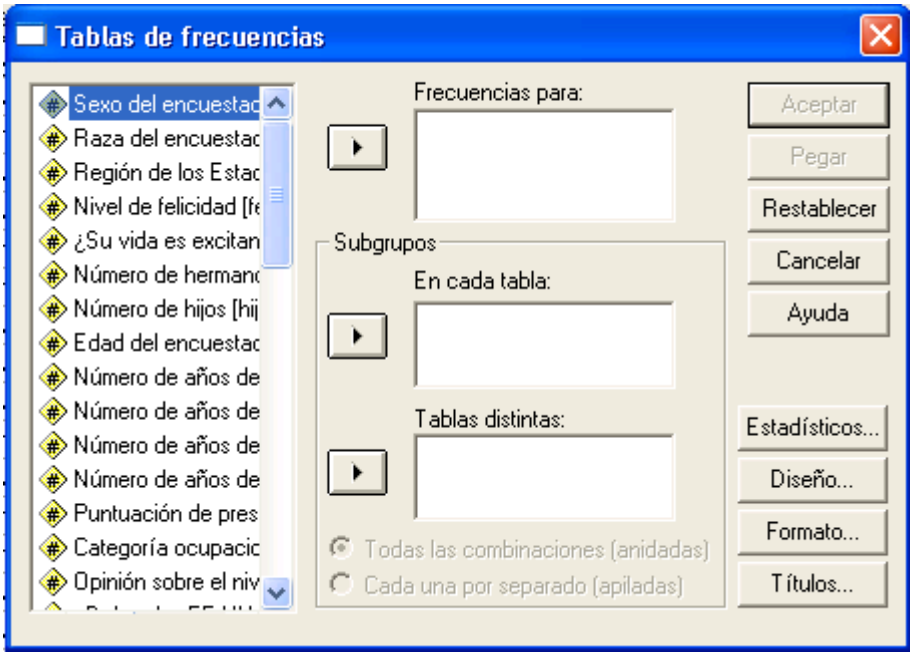


Figura A2.8(b) Cuadro de diálogo de Totales de Tablas básicas

A 2.5 Tablas de frecuencia

Cuando algunas variables comparten la misma categorización, la tablas de frecuencia son una alternativa para presentar dicha información de una forma compacta. El cuadro de diálogo es el de la Figura A2.9.

En el archivo Encuesta general..., que estamos utilizando, hay varias variables (ítems) con los mismas categorías de respuesta. Si pasamos estas variables (obedecer; popular; pensprop; trabajar) a la lista **Frecuencias para** del cuadro de diálogo, el resultado es el siguiente:



	Obedecer es	Ser apreciado y popular es	Pensar por uno mismo es	Trabajar duro es
	Recuento	Recuento	Recuento	Recuento
Lo más importante	195	4	510	147
Lo 2º más importante	123	27	161	355
Lo 3º más importante	142	57	130	321
Lo 4º más importante	343	185	135	144
Poco importante	179	709	46	15

Figura A 2.9 Cuadro de diálogo de Tablas de frecuencias

En este tipo de tablas, los estadísticos son diferentes a los de las tablas básicas, en el sentido de que sólo se pueden solicitar recuentos y porcentajes. El cuadro de diálogo para ello es el de la Figura A 2.10.

Figura A 2.10 Cuadro de diálogo de Estadísticos de Tablas de frecuencia

Se observa en el cuadro de diálogo, que se puede variar el formato de visualización y las etiquetas que aparecen encima de los estadísticos.

A 2.5.1 Añadiendo subgrupos

En su forma más básica, las tablas de frecuencia muestran la distribución de las variables seleccionadas. Sin embargo, se puede ampliar esta información, cruzando estas variables con otras, de modo que la información se pueda visualizar para cada categoría de las variables con las que cruzamos las variables de frecuencia. Así, por ejemplo, si dos de las variables anteriores las cruzamos con la región de procedencia, manteniendo los porcentajes, el resultado es el que se muestra en la tabla siguiente:

	Región de los Estados Unidos									
	Nor-Este				Sur-Este				O	
	Obedecer es		Ser apreciado y popular es		Obedecer es		Ser apreciado y popular es		Obedecer es	
	Recuento	%	Recuento	%	Recuento	%	Recuento	%	Recuento	%
Lo más importante	77	17,07%	1	,22%	68	25,86%	1	,38%	50	18,66%
Lo 2º más importante	55	12,20%	12	2,66%	40	15,21%	9	3,42%	28	10,45%
Lo 3º más importante	65	14,41%	25	5,54%	39	14,83%	10	3,80%	38	14,18%
Lo 4º más importante	169	37,47%	87	19,29%	77	29,28%	47	17,87%	97	36,19%
Poco importante	85	18,85%	326	72,28%	39	14,83%	196	74,52%	55	20,52%
Total	451	100,00%	451	100,00%	263	100,00%	263	100,00%	268	100,00%

Tabla de frecuencia con los subgrupos de la variable Región

La información que muestra esta tabla es mucho más rica que aquella en la que sólo están las variables con las mismas categorías de respuesta, lo cual permite un análisis más detallado de los datos. Obviamente, la tabla se puede complicar mucho si, además de la variable de grupo incorporada en la lista **En cada tabla**, añadimos alguna variable en la lista **Tablas distintas**. Al igual que en las Tablas básicas, si se añade más de una variable en estas dos listas se puede elegir entre el anidamiento o el apilamiento.

A 2.6 Tablas generales

Este procedimiento es el más completo de todos, pues con el se pueden conjugar tanto el anidamiento como el apilamiento de variables, la incorporación de estadísticos diferentes para cada variable, totales más complejos que en las tablas básicas y, además, se puede manejar variables de respuestas múltiples, muy comunes en muchos cuestionarios.

El cuadro de diálogo de este procedimiento es el que se muestra en la Figura A2.11. Al igual que en los demás tipos de tablas, la más sencilla es la de un dimensión, y no merece comentarios. Si se añade más de una variable en la misma dimensión es posible, como ya se ha dicho anidar o apilar las variables. Por defecto, cada variable que se incorpora en la misma dimensión se apila al resto de variables. Si se quiere anidar, es preciso marcar la variable que quiere anidarse a la inmediatamente anterior en la lista y pulsar el correspondiente botón **> Anidar**.

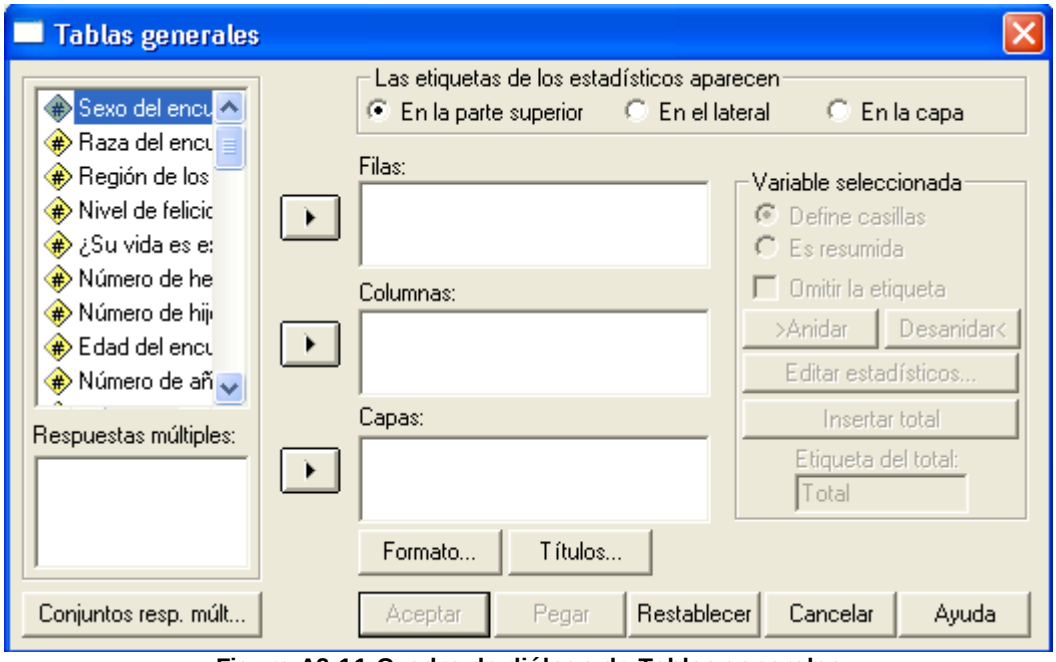
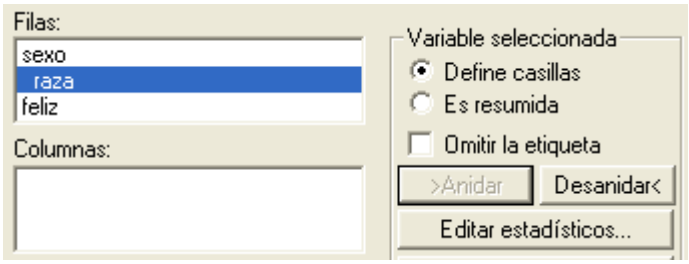


Figura A2.11 Cuadro de diálogo de Tablas generales

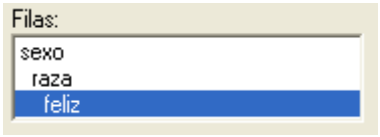
Si introducimos tres variables en una dimensión y apilamos sólo la intermedia, tal como se muestra en la Figura A2.12, el resultado es el siguiente:



Sexo del encuestado	Hombre	Raza del encuestado	Blanca	545
			Negra	71
			Otra	20
	Mujer	Raza del encuestado	Blanca	719
			Negra	133
			Otra	29
Nivel de felicidad	Muy feliz			467
	Bastante feliz			872
	No demasiado feliz			165

Figura A2.12 Ejemplo de tres variables, la intermedia anidada. en una dimensión

También se puede establecer una anidación múltiple, tal como se muestra en la Figura A2.13.



Sexo del encuestado	Hombre	Raza del encuestado	Blanca	Nivel de felicidad	Muy feliz	180
					Bastante feliz	326
					No demasiado feliz	37
			Negra	Nivel de felicidad	Muy feliz	19
					Bastante feliz	38
					No demasiado feliz	13
			Otra	Nivel de felicidad	Muy feliz	7
					Bastante feliz	10
					No demasiado feliz	3
	Mujer	Raza del encuestado	Blanca	Nivel de felicidad	Muy feliz	229
					Bastante feliz	404
					No demasiado feliz	80
			Negra	Nivel de felicidad	Muy feliz	27
					Bastante feliz	78
					No demasiado feliz	26
			Otra	Nivel de felicidad	Muy feliz	5
					Bastante feliz	16
					No demasiado feliz	6

Figura A2.13 Tres variables completamente anidadas en la misma dimensión

A 2.6.1 Añadiendo estadísticos

A diferencia con las tablas básicas, en este procedimiento se puede elegir estadísticos para cada variable, con ciertas limitaciones. Para ello, simplemente se marca la variable una vez que está incorporada a la dimensión que deseemos. El cuadro de diálogo de estadísticos, al que se accede pulsando el botón **Editar estadísticos**, sólo para variables categóricas u ordinales con pocas categorías, es el que se muestra en la Figura A2.14.

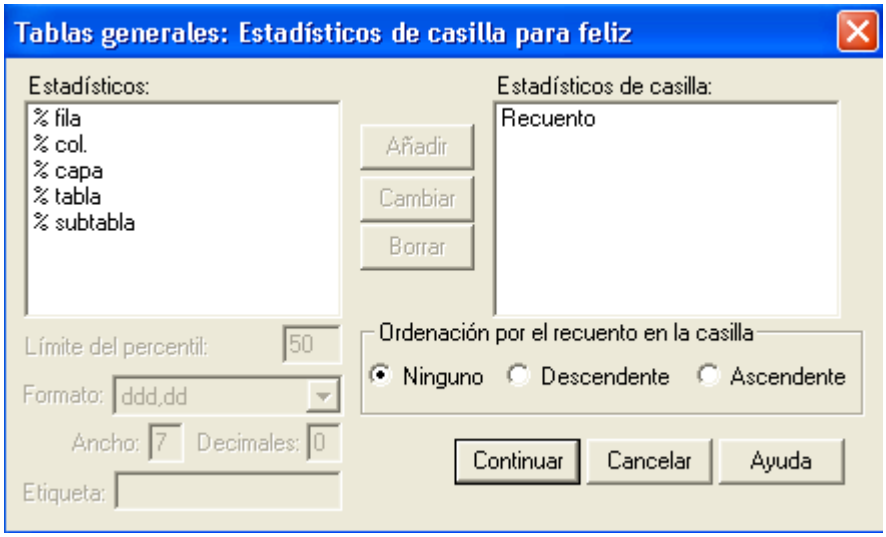


Figura A2.14 Cuadro de diálogo de Estadísticos de casilla para variables categóricas

Cuando la variable tiene un nivel de medida de escala, se puede elegir otros estadísticos, pulsando el mismo botón, y marcando la opción **Es resumida** en el apartado **Variable seleccionada** del cuadro de diálogo principal. Entonces el cuadro de estadísticos es el que se muestra en la Figura A2.15.

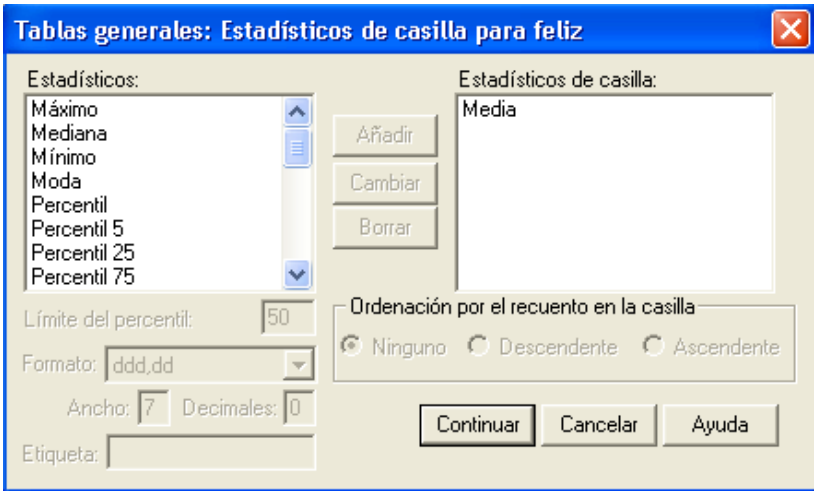


Figura A2.15 Cuadro de diálogo de Estadísticos de casilla para variables de escala

Como ejemplo, se puede ver la siguiente tabla, en la que se han incorporado en la misma dimensión dos variables, una categórica (sexo) y otra de escala (años de educación). Para la primera se ha solicitado el porcentaje sobre la tabla y para la segunda la media y la desviación típica.

			Raza del encuestado		
			Blanca	Negra	Otra
Sexo del encuestado	Hombre	Recuento	545	71	20
		% tabla	35,9%	4,7%	1,3%
	Mujer	Recuento	719	133	29
		% tabla	47,4%	8,8%	1,9%
Número de años de escolarización	Media		(13,06)	(11,89)	(12,47)
	Desviación típ.		(2,95)	(2,68)	(4,00)

Una limitación obvia respecto de los estadísticos, cuando las variables se anidan completamente es que sólo se pueden pedir para la del último nivel de anidamiento; las de los niveles superiores actúan como variables de agrupamiento.

Un último apunte respecto de las variables de escala es que cuando se especifican como variable seleccionada que está resumida, a la derecha de la variable aparece la letra S entre paréntesis tal como puede verse en la Figura A2.16. También se puede ver en la misma Figura cómo una vez que se ha marcado una variable como resumen se muestra esta condición (**Dimensión resumen**) a la derecha de la dimensión en que se ha incorporado.

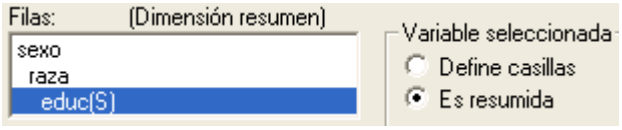


Figura A2.16 Ejemplo de variable marcada como resumida

A 2.6.2 Los totales en las tablas generales

En este tipo de tablas los totales son más flexibles que en la básicas, y es precio especificarlo para cada variable que se desee. Para ello sólo hay que marcar la variable una vez que está en la dimensión correspondiente y pulsar el botón Insertar total. Si, por ejemplo, pedimos totales para dos variables cada una en una dimensión en el cuadro de diálogo se ve los que muestra la Figura A2.17.

		Región de los Estados Unidos			Total
		Nor-Este	Sur-Este	Oeste	
Sexo del encuestado	Hombre	281	177	178	636
	Mujer	398	238	245	881
Total		679	415	423	1517

Figura A2.17. Ejemplo de inserción de totales para variables

Por defecto, un total dará el mismo estadístico que la variable totalizada. Se puede totalizar cualquier variable de la tabla excepto otra total, una variable de resumen o una variable anidada bajo una variable totalizada. Es decir, si en una misma dimensión incluimos, anidadas, sexo y región, podemos totalizar una u otra, pero no ambas a la vez. En las Figuras A2.18(a) y A2.18(b) se muestran sendos ejemplos de inserción de totales.

Figura A2.18(a)

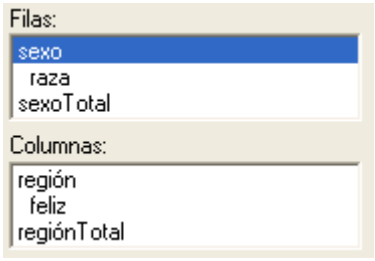
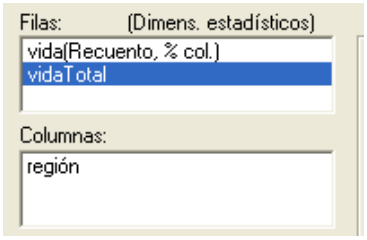


Figura A2.18(b)

Cuando se incorporan variables resumen y se piden totales, el resultado es el que se muestra en la siguiente tabla:

				Región de los Estados Unidos			Total
				Nor-Este	Sur-Este	Oeste	
Sexo del encuestado	Hombre	Edad del encuestado	Media	(43,96)	(45,68)	(43,04)	(44,18)
		Número de años de	Media	(13,30)	(13,11)	(13,26)	(13,23)
	Mujer	Edad del encuestado	Media	(46,98)	(48,95)	(43,98)	(46,67)
		Número de años de	Media	(12,79)	(11,98)	(13,00)	(12,63)
Total	Edad del encuestado	Media		(45,72)	(47,55)	(43,59)	(45,63)
	Número de años de	Media		(13,00)	(12,46)	(13,11)	(12,88)

Se puede dar el caso en que se desee un estadístico para un total diferente al de la variable que totaliza. Un ejemplo se puede ver en la Figura A2.19.



		Región de los Estados Unidos					
		Nor-Este		Sur-Este		Oeste	
		Recuento	% col.	Recuento	% col.	Recuento	% col.
¿Su vida es excitante o aburrida?	Excitante	186	42,96%	107	40,07%	141	50,36%
	Rutinaria	228	52,66%	148	55,43%	129	46,07%
	Aburrida	19	4,39%	12	4,49%	10	3,57%
Total		433	100,00%	267	100,00%	280	100,00%

Figura A2.19 Total por defecto

Si se quiere obtener recuentos del total, es preciso marcar, en este caso, *vida Total* y pulsar **Editar estadísticos**. En el cuadro de diálogo que se abre (Figura A2.20), para que se activen los estadísticos, hay que marcar la opción **Estadísticos del total personalizados**, y elegir el estadístico adecuado.

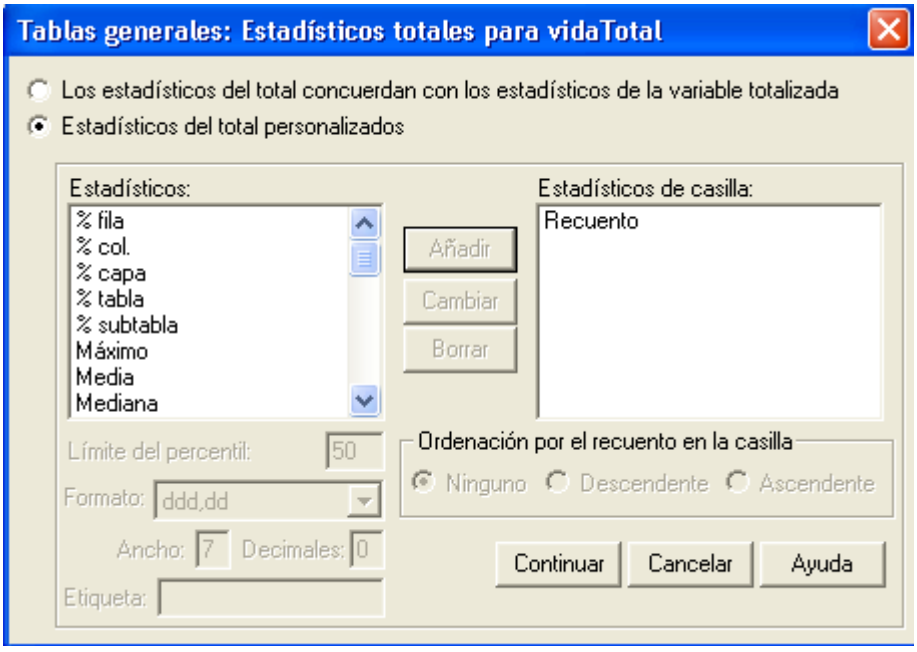
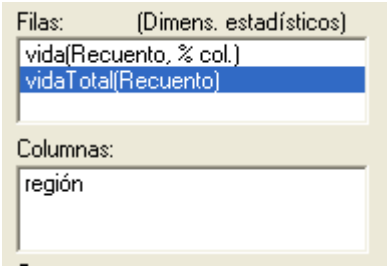


Figura A2.20. Cuadro de diálogo de Estadísticos totales

Al haber elegido estadísticos distintos para variables diferentes en las filas, las etiquetas de las casillas aparecen en las filas, de modo automático, tal como se aprecia en la Figura A2.21.



			Región de los Estados Unidos		
			Nor-Este	Sur-Este	Oeste
¿Su vida es excitante o aburrida?	Excitante	Recuento	186	107	141
		% col.	42,96%	40,07%	50,36%
	Rutinaria	Recuento	228	148	129
		% col.	52,66%	55,43%	46,07%
	Aburrida	Recuento	19	12	10
		% col.	4,39%	4,49%	3,57%
Total	Recuento		433	267	280

Figura A2.21. Estadísticos personalizados para el total

A2.6.3 Los totales globales

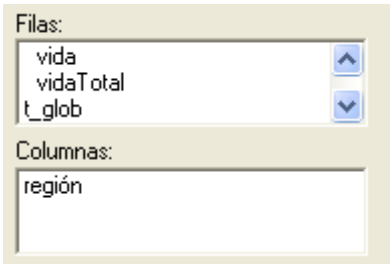
Se puede crear un total de serie de totales, pero para ello es necesario crear una variable auxiliar, que contengo el valor 1 para todos los casos válidos. Para hacer que esta variable auxiliar sea válida sólo para los casos válidos, habrá que especificar dicha condición (que el caso no sea perdido), como en

~ MISSING(vida)

Si la variable está anidada, hay que colocar la operador lógico Y (&) y especificar lo mismo para la variable del nivel superior. Si por ejemplo se anida vida en sexo, la condición será

~ MISSING(vida) & ~ MISSING(sexo)

Cuando ya se haya creado la variable se etiqueta en el Editor de datos el valor 1 como *Total global*, y ya se puede utilizar para crear el total de una serie de totales. Como ejemplo, en la Figura A2.22, se muestra la tabla que se obtiene con las variables *sexo* y *vida* anidadas en una dimensión, en la que, además, se ha incluido la nueva variable generada *t_glob*, cruzadas con la variable *región*.



				Región de los Estados Unidos		
				Nor-Este	Sur-Este	Oeste
Sexo del encuestado	Hombre	¿Su vida es excitante o aburrida?	Excitante	92	56	65
			Rutinaria	88	58	54
			Aburrida	7	3	2
		Total		187	117	121
	Mujer	¿Su vida es excitante o aburrida?	Excitante	94	51	76
			Rutinaria	140	90	75
			Aburrida	12	9	8
		Total		246	150	159
Total global				433	267	280

Figura A2.22 Tabla con un total de totales

A2.7 Preguntas de respuesta múltiple

En muchas ocasiones, los cuestionarios que se administran a los encuestados contienen preguntas a las que se puede dar múltiples respuestas. Se pregunta, por ejemplo, qué diarios leen; qué programas de TV ven habitualmente; o qué características valoran de un líder político. Lógicamente, una sola variable no puede registrar esta diversidad de respuestas, por lo que es preciso generar tantas variables como diferentes respuestas haya estipuladas en el cuestionario de una pregunta concreta.

Como ejemplo de este tipo de preguntas, la Encuesta general USA contiene algunas de este tipo. En primer lugar se pregunta a los encuestados una pregunta abierta a la que pueden dar múltiples respuestas, y posteriormente se les hacen una serie de preguntas específicas a las que sólo puede responder 'sí' o 'no'. Según cuál sea el tipo de pregunta la codificación diferirá en cada caso.

Si, por ejemplo, se plantea la pregunta ¿Cuáles son los problemas más importantes que han tenido en su familia durante el último año?, las respuestas se registran con las palabras exactas del encuestado y luego se codifican para el análisis. Dicha respuesta amplia se puede dividir en varias categorías cada una con otra serie de categorías a su vez. Si la respuesta contiene un problema se codifica en la primera variable; si tiene dos, se codifican en la primera y segunda variables, y así sucesivamente. Esta forma de codificar se denomina **respuestas múltiples codificadas como categorías**.

Hay también otro tipo de preguntas, más específicas, relacionadas, por ejemplo, con asuntos de salud, encabezadas todas ellas con la pregunta genérica ¿se ha encontrado durante el año pasado con alguna de estas situaciones?

1. Pedir consejos por problemas matrimoniales o emocionales.
2. Padecer de infertilidad, o no poder tener hijos.
3. Consumo de drogas en general.
4. Muerte de un amigo próximo…

Cada una de estas preguntas se responden con 'sí' o 'no/no respuesta/no aplicable'. Este conjunto de respuestas se denominan **pregunta de respuesta múltiple codificadas como dicotomías**.

A2.7.1 Definición de conjuntos de respuestas múltiples

Para que se puedan utilizar variables de respuesta múltiple en una tabla, previamente hay que definirlas. Para ello, se pulsa el botón **Conjuntos de respuestas múltiples**, en el cuadro de diálogo de Tablas generales y se accede al cuadro de diálogo de la Figura A2.23.

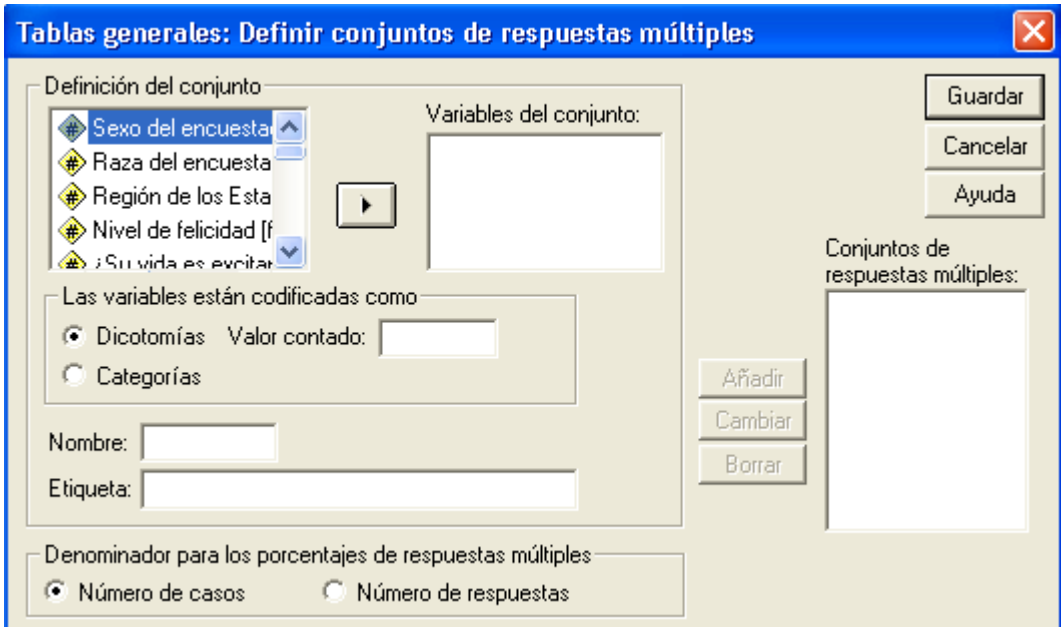


Figura A2.23 Cuadro de diálogo para definir conjuntos de respuestas múltiples

A 2.7.1.2 Definición de conjuntos como categorías

Para generar un conjunto de variables codificadas como categorías, primero se seleccionar las variables que lo van a formar y se desplazan a la lista **Variables del conjunto**. Si las variables de respuesta múltiple se codifican como categorías, es preciso señalarlo en el apartado correspondiente del cuadro de diálogo. Luego se le da nombre y si se desea una etiqueta que explique en qué consiste dicho conjunto. Por último, se pulsa el botón **Añadir** y se incorpora a la lista **Conjuntos de respuesta múltiple**.

En el archivo de datos hay una serie de variables (prob1 a prob4) que son variables multicategóricas. Para generar un conjunto, seleccionamos estas variables como integrantes de un conjunto, al que damos el nombre **probcat**, y como etiqueta '*Problemas más significativos en último añ*o'. El conjunto generado tendrá como nombre **$probcat**. y en el cuadro de diálogo general aparece tal como se muestra en la Figura A2.24.

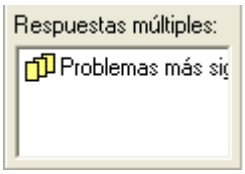


Figura A2.24 Conjunto de respuesta múltiple ya generado

Este conjunto ya se puede utilizar como una variable más dentro del procedimiento Tablas generales. Si, por ejemplo, utilizamos el conjunto en la dimensión de las filas y le insertamos el total, el resultado es el de la Figura A2.25.

Problemas más signficativos en último año	Salud	138
	Dinero	208
	Falta de servicios básicos	7
	Familia	76
	Motivos personales	41
	Problemas legales	2
	Varios	71
Total		336

Figura A2.25 Tablas de un conjunto de respuestas múltiple de variables multicategóricas

A 2.7.1.3 Definición de conjuntos como dicotomías

El procedimiento es similar al de las variables categóricas, pero al ser variables dicotómicas (respuestas de sí o no) es preciso señalar, cuál es el valor que se ha utilizado para designar la respuesta 'sí'. En el archivo Encuesta…, hay un total de 18 preguntas de este tipo, referidas a salud y trabajo (9 de cada). Si generamos un conjunto con estas 18 variables, y luego lo utilizamos en la dimensión de las filas el resultado es el que se muestra en la Figura A2.26

		Casos
Problemas de salud y trabajo	Suficientemente enfermo para ir al médico	559
	En tratamiento psicológico	58
	Esterilidad, incapaz de tener hijos	35
	Problemas con el alcohol	17
	Drogas ilegales (Marihuana, Cocaína)	30
	Cónyugue en el hospital	73
	Hijo en el hospital	78
	Hijo drogadicto o alcohólico	28
	Muerte de un amigo próximo	230
	Parado o buscando empleo más de un mes	59
	Degradado o relegado a una posición peor	26
	Reducción en el salario o en el horario	61
	Relegado en la promoción	38
	Problemas con el jefe	42
	Negocio propio con problemas o perdiendo dinero	20
	Cónyugue despedido	25
	Reducción salarial al cónyugue	52
	La esposa está desempleada	55
Total		770

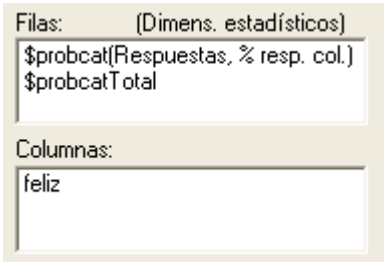
Figura A2.26 Tabla de un conjunto de respuesta múltiple de variables dicotómicas

A 2.7.2 Uso de conjuntos de respuesta múltiple

Como ya hemos señalado. los conjuntos de respuesta múltiple se pueden tratar como variables categóricas normales. Así por ejemplo, si cruzamos el conjunto **$probcat** con la región de residencia, el resultado es el siguiente

		Región de los Estados Unidos		
		Nor-Este	Sur-Este	Oeste
Problemas más signficativos en último año	Salud	68	40	30
	Dinero	98	48	62
	Falta de servicios básicos	4	1	2
	Familia	33	15	28
	Motivos personales	19	9	13
	Problemas legales	1		1
	Varios	29	18	24
Total		158	83	95

Un aspecto importante respecto de las tablas con conjuntos de respuesta múltiple es que el número de respuestas puede ser mayor que el número de casos, dado que un mismo sujeto puede ofrecer más de una respuesta dentro de la misma pregunta. Por ello, son dos las formas de calcular porcentajes: o bien todos los casos suman el 100%, o bien todas las respuestas suman el 100%. Por defecto todos los casos suman el 100%, tal como puede verse en la Figura A2.24.



		Nivel de felicidad					
		Muy feliz		Bastante feliz		No demasiado feliz	
		Respuestas	% resp. col.	Respuestas	% resp. col.	Respuestas	% resp. col.
Problemas más notables en último año	Salud	38	47,5%	75	37,1%	23	47,9%
	Dinero	37	46,3%	130	64,4%	36	75,0%
	Falta de servicios básicos	2	2,5%	5	2,5%		
	Familia	15	18,8%	44	21,8%	16	33,3%
	Motivos personales	11	13,8%	24	11,9%	6	12,5%
	Problemas legales	1	1,3%	1	,5%		
	Varios	17	21,3%	43	21,3%	11	22,9%
Total		121	151,3%	322	159,4%	92	191,7%

Figura A2.27 Casos que suman más del 100%

En la tabla se observa que los totales de columna son superiores al 100%, dado que hay más repuestas que casos. Si se desea ajustar el porcentaje a las

respuestas y no a los casos, hay que marcar la opción en el cuadro de diálogo de definición de conjuntos de respuesta múltiple, tal como puede verse en la Figura A2.28.

		Nivel de felicidad					
		Muy feliz		Bastante feliz		No demasiado feliz	
		Respuestas	% resp. col.	Respuestas	% resp. col.	Respuestas	% resp. col.
Problemas más notables en último año	Salud	38	31,4%	75	23,3%	23	25,0%
	Dinero	37	30,6%	130	40,4%	36	39,1%
	Falta de servicios básicos	2	1,7%	5	1,6%		
	Familia	15	12,4%	44	13,7%	16	17,4%
	Motivos personales	11	9,1%	24	7,5%	6	6,5%
	Problemas legales	1	,8%	1	,3%		
	Varios	17	14,0%	43	13,4%	11	12,0%
Total		121	100,0%	322	100,0%	92	100,0%

Figura A2.28 Respuestas que suman el 100%

También se pueden combinar en la misma tabla el número de casos y el número de respuestas siendo el resultado el que se muestra en la tabla siguiente:

Apéndice 2. El módulo de Tablas

		Nivel de felicidad					
		Muy feliz		Bastante feliz		No demasiado feliz	
		Casos	Respuestas	Casos	Respuestas	Casos	Respuestas
Problemas más notables en último año	Salud	38	38	75	75	23	23
	Dinero	37	37	130	130	36	36
	Falta de servicios básicos	2	2	5	5		
	Familia	15	15	44	44	16	16
	Motivos personales	11	11	24	24	6	6
	Problemas legales	1	1	1	1		
	Varios	17	17	43	43	11	11
Total		80	121	202	322	48	92

Aunque los valores para cada categoría de la variable **$probcat** presentan los mismos valores, el total de cada uno es diferente. El total de respuestas muestra el total de respuesta, mientras que el total de casos muestra el total de casos.

Bibliografía

Agresti, A. (1990). *Categorical data analysis.* Nueva York: Wiley.

Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistic Society,* 9, 176- 197.

Bock, R. D. (1975). *Multivariate Statistical methods in behavioral research.* Nueva York: McGraw-Hill.

Box, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics,* 25, 290-302.

Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics,* 25, 484-498.

Box, G., Hunter, W., y Hunter, J. (1989). *Estadística para investigadores. Introducción al diseño de experimentos, análisis de datos y construcción de modelos*. Ed. Reverté. Barcelona.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika,* 37, 256-266.

Cochran, W. G. (1952). The X^2 test of goodness of fit. *Annals of Mathematical Statistics,* 23, 315-345.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20,37-46.

Conover, W. J. (1980). *Practical nonparametric Statistics* (2.. ed.). Nueva York: Wiley. Cook, R. D. (1977) .Detection of influential observations in linear regression. *Technometrics,*19, 15-18.

Cramer, H. (1946). *Mathematical methods of Statistics.* Princeton, NJ: Princeton University Press.

Duncan, D. B. (1955). Mu1tiple range and multiple F tests. Biometrics, 11, 1-42.

Dunn, C. W. (1961). Multiple comparisons among means. *Journal of the American Statistical Association,* 56, 52-64.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal *of the American Statistical Association,* 50, 1096-1121.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association,* 75,795-800.

Durbin, J. y Watson, G. S. (1951). Testing for serial correlation in least-squares regression II. *Biometrika,38,* 159-178

Fisher, R. A. (1935). *Statistical methods for research workers* (5º ed.). Edinburgo: 0liver and Boyd (14ª. ed. en 1973: Nueva York: Hafner).

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.

Friedman, M. (1937). The use ofranks to avoid the assumption ofnormality implicitin the analysis of variance. *Journal of the American Statistical Association,* 61, 1081-1096.

Bibliografía

Gabriel, K. R. (1969). Simultaneous test procedures: Some theory of multiple comparisons. *Annals of Mathematical Statistics,* 40, 224-240.

Games, P. A. y Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal *n*'s and/or variances: A MonteCarlo study*. Journal of Educational Statistics,* 1,113-125.

Geisser, S. y Greenhouse, S. W. (1958). An extension ofBox' results on the use *of F* distribtion in multivariate analysis. *Annals of Mathematical Statistics,* 29,885-891.

Goodman, L. A. y Kruskal, W. H. (1979). *Measures of associationfor cross classiflcations.* Nueva York: Springer-Verlag.

Haberman, S. J. (1973). The analysis of residuals in cross-classification tables. *Biometrics,* 29, 205-220.

Huynh, H. y Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot design. *Journal of Educational Statistics,* 1, 69-82.

Kendall, M. G. (1963). *Rank correlation methods* (3.. ed.). Londres: Griffin (4ª. ed. en 1970).

Kendall, M. G. y Babington-Smith, B. (1939). The problem of *m* rankings. The Annals of Mathematical Statistics, *10,* 275-287.

Keppel, G. (1982). *Design and analysis: a research's handbook*. Englewood Cliffs, New Jersey : Prentice-Hall, 1982

Kirk, R. E. (1982). *Experimental design. Procedures for the behavioral sciences* (2ª ed.). Belmont, CA: Brooks/Cole (3ª ed. en 1995).

Kruskal, W. H. y Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis. *Journal of the American Statistical Association,* 47, 583 -621

Levene, H. (1960). Robust tests for the equality of variances. En J. Olkin (Ed.): *Contributions to probability and statistics.* Palo Alto, CA: Stanford University Press.

Lillieffors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and va- riance unknown. *Journal of the American Statistical Association,* 62,399-402.

Mahalanobis, P. C.(1936). On the generalized distance in statistics. *Procedures National Science India,* 2,49-55.

Mann, H. B. y Whitney, D. R. (1947). Ona test of whetherone of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18, 50-60.*

Mantel,N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute,* 22, 719- 748.

Marascuilo, L. A. y McSweeney, M. (1977). *Nonparametric and distribution-free methods* Monterrey, CA: Brooks/Cole.

Mauchly, J. W. (1940) .Significance test for sphericity of a normal *n*- variate distribution. *Annals of Mathematical Statistics, II,* 204-209.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12,* 153-157.

Moses, (1952). A two sample test. *Psychometrika, 17,239-247.*

Norusis, M. J. y SPSS, Inc. (1993). *SPSS for Windows. Base system user's guide release* 6.0. Chicago, IL: SPSS Inc.

Pardo, A. (2002). *Análisis de datos categóricos*. UNED Ediciones. Madrid

Scheffé, H. A. (1953). A method for judging all possible contrasts in the analysis of variance. *Biometrjka,* 40,87-104.

Scheffé, H. A. (1959). *The analysis of varjance.* Nueva York: Wiley.

Shapiro, S. S. y Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrjka,* 52,591-611.

Smimov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow Universjty,* 2,3-16 [ruso].

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review,* 27,799-811.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology,* 15,72-101.

SPSS (1991). *SPSS statistjcal algorjthms* (2.. ed.). Chicago, IL: SPSS Inc.

Tukey, J. W. *(1977).Exploratory data analysis.* Reading: Addison Wesley.

Wald, A. y Wolfowitz, J. (1940). On a test whether two samp1es are from the same population. *Annals of Mathematical Statistics,* 11,147-162.

Waller, R. A. y Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparison problem. *Journal of the American Statistical Association,* 64, 1484-1503.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika.* 29,350-362.

Welsch,R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association,* 72,566-575.

Wilcoxon, F. (1945).lndividual comparisons by ranking methods. *Biometrics,* 1, 80-83. Wilcoxon, F. (1949). *Some rapid approximate statistical procedures.* American Cyanamid Co., Standford Research Laboratories.

Winer, Brown y Michels (1991). *Statistical principles in experimental design* (3.. ed.). Nueva York: McGraw-Hill.

Yates, F. (1934). Contingency tables involving small numbers and the X^2 test. *Journal of the Royal Statistical Society, supplement* 1,217-235.

1ª Edición: Madrid, octubre de 2008