Lic. Nel Quezada Lucio

# Estadística con SPSS 22

- Investigación y manejo de datos
- Manipulación de casos y variables
- Informes, estadísticas descriptivas y tablas
- Pruebas no paramétricas
- Gráficos




EMPRESA EDITORA MACRO®

Incluye CD Multimedia

EDITORIAL MACRO®

# Estadística con SPSS 22



**Estadística con SPSS 22**
Autor: Nel Quezada Lucio

© Derecho de autor registrados:
Empresa Editora Macro E.I.R.L.

© Derecho de edición, arte gráfico y diagramación reservados
Empresa Editora Macro E.I.R.L.

**Corrector de estilo:**
Jorge Giraldo Sánchez

**Coordinador de arte y diseño:**
Alejandro Marcas León

**Diagramación:**
Cynthia Arestegui Baca

Edición a cargo de:
Empresa Editora Macro E.I.R.L.
Av. Paseo de la República N° 5613 , Miraflores, Lima, Perú

(511) 748 0560
ventas@editorialmacro.com
www.editorialmacro.com

Primera edición: junio 2014
Tiraje 1 000 ejemplares

Impresión:
Talleres Gráficos de la Empresa Editora Macro E.I.R.L.
Jr. San Agustín N.° 612-624, Surquillo, Lima, Perú
Lima, Perú

ISBN Nº 978-612-304-206-6
Hecho el Depósito Legal en la Biblioteca Nacional del Perú Nº 2014-06513

Prohibida la reproducción parcial o total, por cualquier medio o método, de este libro sin previa autorización de la Empresa Editora Macro EIRL.

## NEL QUEZADA LUCIO

Licenciado en Estadística por la Universidad Nacional de Ingeniería (UNI). Candidato a Magíster en Ciencias con mención en Estadística Matemática en la Universidad Nacional Mayor de San Marcos (UNMSM). Posee una maestría en Ciencias, con mención en Ingeniería de Sistemas-UNI. Es catedrático de la Escuela Profesional de Ingeniería y Estadística de la Universidad Nacional de Ingeniería. Además, es miembro del directorio de Cerveza Taski, asesor y consultor en diversas empresas.

Ha realizado trabajos en instituciones públicas y privadas: diseño y desarrollo de productos con estudios de *Panel de consumidores*, *Tamaño de mercado y segmentación de mercado*; en *Sistema de información estadística*, así como en la *Elaboración, crítica, codificación, procesamiento, control de calidad y análisis*; en el desarrollo y gestión de estudios e investigaciones en diferentes campos; en teoría de muestreo aplicado en encuestas psicológicas y socioeconómicas para efectos de evaluar el comportamiento de la población.

## Dedicatoria

A Lucas Nicolás Quezada Meza.

# Índice


INTRODUCCIÓN ..... 11

**CAPÍTULO 1: INTRODUCCIÓN A SPSS ..... 13**

1.1 SPSS 22 ..... 15
1.2 INICIAR SPSS 22 ..... 15
1.2.1 Pasos básicos a seguir ..... 16
1.2.2 Barra de menú ..... 19
1.2.3 Barra de herramientas ..... 21
1.2.4 Ventanas ..... 21

**CAPÍTULO 2: MANEJO DE DATOS ..... 25**

2.1. INTRODUCCIÓN DIRECTA DE DATOS ..... 27
2.1.1 Realizar la encuesta ..... 28
2.1.2 Codificar la encuesta ..... 28
2.1.3 Llenar la base de datos (Editor de datos) ..... 30
2.2. EXTRAER UN ARCHIVO DE DISCO ..... 35
2.2.1 Un paseo por el menú ..... 36
2.2.2 Uso de los cuadros de diálogo ..... 47

**CAPÍTULO 3: MANIPULACIÓN DE CASOS Y VARIABLES ..... 57**

3.1 VER INFORMACIÓN DE DATOS (DISPLAY DATA INFO…) ..... 59
3.2 VARIABLES ..... 60
3.3 BUSCAR DATOS… (FIND) ..... 60
3.4 IR A CASO ..... 61
3.5 INSERTAR VARIABLE (DATA > INSERT VARIABLE…) ..... 62
3.6 INSERTAR CASO (DATA > INSERT CASE…) ..... 62
3.7 ORDENAR CASOS (DATA > SORT CASES…) ..... 63
3.8 DIVIDIR ARCHIVO (DATA > SPLIT FILE…) ..... 63
3.9 SELECCIONAR CASOS (DATA > SELECT CASES…) ..... 64
3.10 CALCULAR VARIABLE (COMPUTE VARIABLE…) ..... 66
3.11 RECODIFICAR EN LAS MISMAS VARIABLES ..... 67
3.12 RECODIFICAR EN DIFERENTE VARIABLE… (INTO DIFFERENT VARIABLE…) ..... 69
3.13 RECODIFICACIÓN AUTOMÁTICA… ..... 70
3.14 VENTANA DE DIÁLOGO HERRAMIENTA MEMORIA (RECALL TOLL) ..... 70
3.15 HERRAMIENTA ESTRUCTURAS DE SINTAXIS ..... 71
3.16 BOTÓN EJECUTAR PROGRAMA ..... 71

**CAPÍTULO 4: INFORMES, ESTADÍSTICAS DESCRIPTIVAS, TABLAS ..... 73**

4.1 INFORMES (REPORTS) ..... 75
4.1.1 Ejercicio práctico 1 ..... 76
4.1.2 Cubos OLAP ..... 76
4.1.3 Resúmenes de casos ..... 79
4.2 ESTADÍSTICOS DESCRIPTIVOS ..... 82
4.2.1 Ejercicio práctico 1 ..... 83
4.2.2 Ejercicio práctico 2 ..... 105
4.3 TABLAS PERSONALIZADAS ..... 111
4.3.1 Ejercicio práctico 1 ..... 112
4.3.2 Ejercicio teórico ..... 116

**CAPÍTULO 5: COMPARAR MEDIAS, MODELO LINEAL GENERAL Y CORRELACIÓN ..... 119**

5.1 COMPARAR MEDIAS (PRUEBA DE HIPÓTESIS) ..... 121
5.1.1 Ejercicio práctico 1 ..... 122
5.1.2 Ejercicios propuestos ..... 136
5.2 MODELO LINEAL GENERAL ..... 138
5.2.1 Conceptos básicos ..... 138
5.2.2 Ejercicio práctico 1 ..... 140
5.2.3 Ejercicios propuestos ..... 146
5.3 CORRELACIÓN ..... 147
5.3.1 Ejercicio práctico 1 ..... 148
5.3.2 Ejercicio práctico 2 ..... 155

**CAPÍTULO 6: REGRESIÓN LINEAL Y LOGLINEAL ..... 159**

6.1 REGRESIÓN ..... 161
6.1.1 Ejercicio práctico 1 ..... 163
6.1.2 Ejercicio práctico 2 ..... 176
6.1.3 Ejercicio propuesto ..... 178
6.2 LOGLINEAL ..... 180
6.2.1 Ejercicio práctico 1 ..... 181

**CAPÍTULO 7: CLASIFICAR, REDUCCIÓN DE DATOS Y ESCALAS ..... 187**

7.1 CLASIFICAR ..... 189
7.1.1 Conglomerados de K medias ..... 189
7.1.2 Conglomerados jerárquicos ..... 189
7.1.3 Discriminante ..... 189

7.2 REDUCCIÓN DE DATOS ..... 210
7.2.1 Análisis factorial ..... 210
7.2.2 Análisis de correspondencias ..... 210
7.2.3 Escalamiento óptimo ..... 211
7.3 ESCALAS ..... 231
7.3.1 Análisis de fiabilidad ..... 231
7.3.2 Escalamiento multidimensional ..... 231
7.3.3 Escalamiento multidimensional (PROXSCAL) ..... 231

**CAPÍTULO 8: PRUEBAS NO PARAMÉTRICAS ..... 233**

8.1 PRUEBAS NO PARAMÉTRICAS ..... 235
8.2 EJERCICIO PRÁCTICO 1 ..... 236
8.3 CONTRASTES NO PARAMÉTRICOS ..... 236
8.3.1 Pruebas de los signos (Binomial) ..... 236
8.3.2 Constrastes para dos muestras ..... 238
8.3.3 Constrastes para más de dos muestras ..... 240
8.4 EJERCICIO PRÁCTICO 2 ..... 243
8.5 PRUEBA DE PEARSON ..... 247
8.5.1 $X^2$ sobre bondad de ajuste ..... 247
8.6 TABLAS DE CONTINGENCIA ..... 249
8.7 EJERCICIOS PROPUESTOS ..... 252

**CAPÍTULO 9: SERIES TEMPORALES, SUPERVIVENCIA Y RESPUESTAS MÚLTIPLES ..... 253**

9.1 SERIES TEMPORALES ..... 255
9.1.1 Ejercicio práctico 1 ..... 256
9.1.2 Series temporales con SPSS ..... 257
9.2 SUPERVIVENCIA ..... 269
9.2.1 Ejercicio práctico 1 ..... 270
9.3 RESPUESTAS MÚLTIPLES ..... 278
9.3.1 Ejercicio práctico 1 ..... 279

**CAPÍTULO 10: GRÁFICOS ..... 285**

10.1 GALERÍA DE GRÁFICOS ..... 287
10.2 SERIE TEMPORAL ..... 297
10.2.1 Serie temporal: Autocorrelaciones ..... 298
10.2.2 Serie temporal: Correlaciones cruzadas ..... 298
10.2.3 Serie temporal: Análisis espectral ..... 298

10.3 COR ..... 299
10.4 GALERÍA DE GRÁFICOS INTERACTIVOS ..... 299
10.4.1 Creación de gráficos interactivos ..... 299
10.4.2 Gráficos de barras interactivos ..... 300
10.4.3 Gráficos de puntos interactivos ..... 301
10.4.4 Gráficos de líneas interactivos ..... 302
10.4.5 Gráficos de bandas interactivos ..... 302
10.4.6 Gráficos de líneas verticales interactivos ..... 303
10.4.7 Gráficos de áreas interactivos ..... 303
10.4.8 Gráficos de sectores simples interactivos ..... 303
10.4.9 Gráficos de sectores agrupados interactivos ..... 304
10.4.10 Gráficos de sectores dispersados interactivos ..... 304
10.5 CREACIÓN DE GRÁFICOS SEGÚN LA ESTRUCTURA DE LOS DATOS ..... 307

**CAPÍTULO 11: PERSONALIZACIÓN SPSS ..... 309**

11.1 CUADRO GENERAL DE OPCIONES ..... 311
11.1.1 Descripción de las opciones ..... 311
11.2 PERSONALIZACIÓN DE LAS BARRAS DE HERRAMIENTAS ..... 317
11.3 PERSONALIZACIÓN DE LAS BARRAS DE MENÚS ..... 318
11.4 PROCESAMIENTO ..... 319
11.5 AUTOPROCESAMIENTO ..... 319

**CAPÍTULO 12: AYUDA ..... 321**

12.1 AYUDA ..... 323
12.2 TEMAS ..... 324
12.3 GUÍA DE SINTAXIS ..... 324
12.4 AYUDA CONTEXTUAL ..... 324
12.5 ASESOR DE RESULTADOS ..... 325
12.5.1 Asesor estadístico ..... 325
12.6 PRODUCCIÓN AUTOMATIZADA ..... 326

**GLOSARIO DE TÉRMINOS** ..... 327

**APÉNDICE** ..... 334

**FUENTES BIBLIOGRÁFICAS** ..... 335
**Referencias electrónicas** ..... 335

# Introducción

Estadística es una ciencia que sirve para tomar decisiones en situaciones de incertidumbre; en otras palabras, es la ciencia autorizada de diseñar, recolectar, describir, analizar e interpretar la información y, por ello, constituye la metodología principal que permite *transformar datos en información*. Esto la hace especialmente atractiva, ya que en cualquier sociedad desarrollada, existe una inmensa variedad de problemas cuya solución solo es posible mediante el uso de técnicas estadísticas. Así, por ejemplo, la estadística es imprescindible, para modelar y predecir diversas variables como económicas, sociológicas, médicas, psicológicas, ambientales o para detectar factores de riesgo asociados a una determinada variable; establecer la eficacia de una variable, en desarrollo de productos, series económicas, tráfico telefónico, etc. He aquí la importancia del software estadístico SPSS 22, una herramienta muy poderosa que facilita los cálculos estadísticos. Por otro lado, las potentes técnicas analíticas del SPSS 22 le ayudan a realizar un análisis más acertado, con el fin de tomar las decisiones más adecuadas para hacer crecer su organización o empresa.

El objetivo del presente libro titulado *Estadística con SPSS 22* es cumplir el compromiso de ayudarle a convertir el análisis predictivo en una parte vital de sus operaciones comerciales o de investigación. En consecuencia, se tratará de conseguir un mayor aprovechamiento de la información con los datos disponibles, mediante casos prácticos de investigación. Se analizarán los datos a través de estadísticas descriptivas, probabilidad, regresión y correlación, modelos de análisis de varianza (ANOVA), pruebas no paramétricas, etc. Se estudiará la influencia de una serie de factores en el comportamiento de una variable de interés; asimismo se modelará la respuesta de una variable por medio de regresión lineal múltiple, componentes principales, análisis factorías, análisis de correspondencia, análisis de discriminante y técnicas de clasificación; siempre se buscará el modelo que mejor se adecúe a los datos.

Al respecto, cabe mencionar que en la actualidad, las numerosas investigaciones se centran en las técnicas estadísticas descritas. Todas estas materias se expondrán con claridad y gran aplicabilidad práctica a lo largo del presente libro, permitiendo resolver problemas de negocios e investigación.
SPSS 22 está dirigido principalmente a empresarios, estudiantes y profesionales de ingeniería, economía, administración, sociología, psicología, medicina, biología, ciencias ambientales, etc. y a todo aquel que tome decisiones, con el fin de aumentar sus posibilidades de éxito en el futuro; además le permitirá conocer lo que sucede en su organización en el presente y adelantarse al futuro, gracias al análisis predictivo.

Respecto a la metodología, se adjunta un CD titulado *Introducción al SPSS 22 (Statistical Package for the Social Sciences)*, donde se presentan las características más relevantes de SPSS 22; además, se verán las herramientas estadísticas frecuentemente empleadas; es decir, se explicará brevemente sobre las bondades de SPSS 22. Asimismo, contiene VIDEOS TUTORIALES en el cual se muestran videos prácticos que le permitirán entender, con claridad, algunos procedimientos de mayor uso de SPSS 22; de este modo, le ayudará a visualizar, entender y aprender rápidamente. Se presenta también ARCHIVOS Y EJERCICIOS donde se exponen los ejercicios prácticos del libro impreso que le permitirán un mejor aprendizaje; y, por último, este CD contiene una lista de los links de las principales páginas donde usted encontrará materiales de consulta especializado en SPSS 22.

CAPÍTULO

1

# INTRODUCCIÓN A SPSS

El objetivo principal del presente capítulo es brindar al lector los conceptos y procedimientos básicos para manipular el software SPSS 22, dentro del entorno Windows. Se verá principalmente la estructura del programa, los módulos más importantes, los pasos básicos a seguir, la barra de menús, las herramientas y las principales ventanas.

## 1.1. SPSS 22

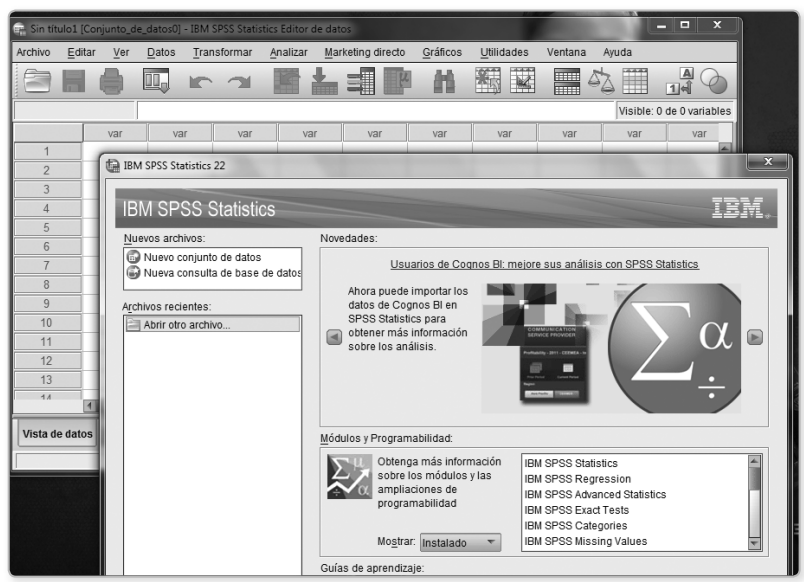
Este software ofrece un rápido entorno de modelación estadístico visual, que va desde lo más simple hasta lo más complejo, para crear modelos de manera interactiva y realizar cambios, utilizando técnicas analíticas probadas y acreditadas. Permite sacar el máximo provecho a los datos, por medio de una completa gama de herramientas estadísticas. Asimismo, ofrece un amplio abanico de posibilidades para llevar a cabo todo el proceso analítico, proporcionándole las respuestas que las hojas de cálculo y las bases de datos no le pueden facilitar. Además, proporciona la información que se necesita para tomar las decisiones más acertadas, usando potentes estadísticos, entendiendo y presentando los resultados en formato de tablas de gran calidad o bien en forma de gráficos, pudiendo compartir los resultados con otras personas por medio de diferentes métodos, incluyendo la publicación de los datos en la Web.

Se trata de un producto modular perfectamente integrado, indicado para todas las etapas del proceso analítico: planificación, recogida de datos, acceso y tratamiento de los mismos, y presentación de los resultados. Constituye el software ideal para resolver los problemas que surgen diariamente en las empresas y organismos oficiales recurriendo al análisis estadístico.

Igualmente, su potente interfaz gráfica es muy sencilla de utilizar. Incluye todos los métodos de análisis y presentación de resultados que usted necesita, para llevar a cabo incluso los análisis más complicados. SPSS 22 es una herramienta para el análisis y administración de datos capaz de producir gráficos y reportes estadísticos rápidos, con una gran variedad de técnicas estadísticas especializadas. Está diseñado en un entorno gráfico de modo que pueda accederse a todos los procedimientos señalando y pulsando con el mouse.

## 1.2. INICIAR SPSS 22

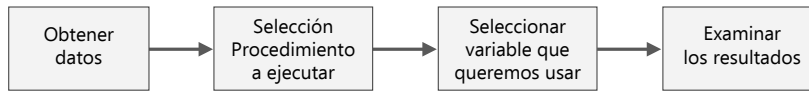
Para inciar el software SPSS 22, primero se accede al botón Inicio, y desde allí a Programas. Luego SPSS 22 for Windows. Se muestra la presentación del programa y, al fondo, el Editor de datos:




Si hace clic en Cancelar y se pasa directamente a la ventana Editor de datos. La otra opción es seleccionar una de las opciones de la ventana Módulos y programatibidad o Guias de aprendizaje, luego haga clic en Aceptar.

### 1.2.1. PASOS BÁSICOS A SEGUIR

Es importate tener presente que se deben desarrollar cuatro pasos básicos para empezar a trabajar con el software SPSS 22, los mismos que se detallarán a continuación:



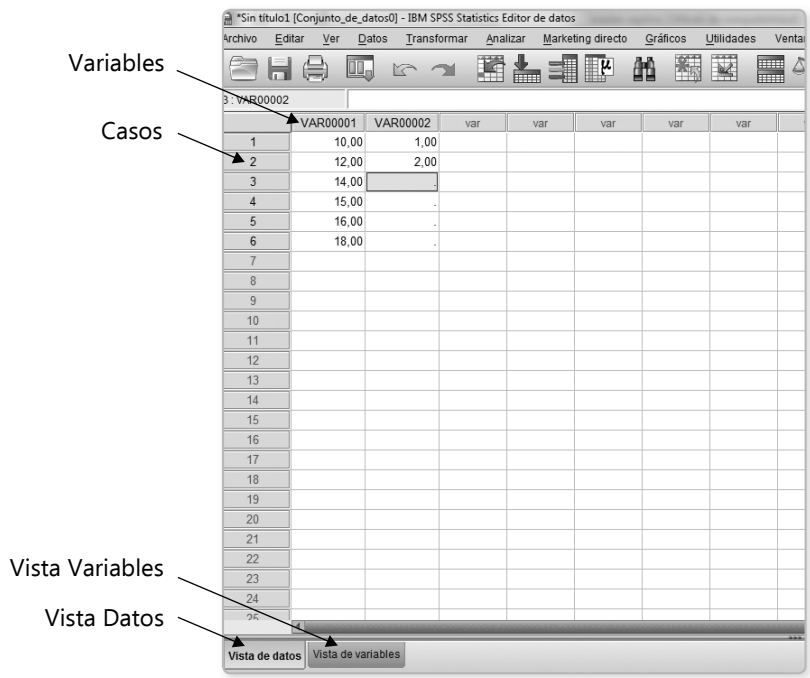

**A. Obtener datos**

La recolección de datos puede realizarse a partir de fuentes existentes o a través de la observación o estudios experimentales. Para obtener datos en SPSS 22, existen dos principales modos que se desarrollarán a continuación:

**a. Introducción directa (observación o estudios experimentales)**

Introducción directa de los datos en las celdas que muestra el Editor de datos (vista Datos). Las columnas indican las variables que están en estudio (VAR00001, VAR00002, VAR00003, VAR00004, VAR00005...); y las filas, los casos de las variables (1, 2, 3, 4, 5, 6, 7...).

Al introducir un dato en una casilla, se activa la variable de la columna donde esté y se defina automáticamente el tipo de dato que se haya introducido. Si quiere definir las variables, debe seleccionar la Vista de variables, donde se podrá especificar el nombre (name), tipo (type), tamaño (width), decimales (decimals), Label (etiqueta) y otros atributos.




**b. Extraer un archivo de disco (fuentes existentes)**

Obtener datos extrayendo un archivo de disco, el cual puede ser de formato SPSS (*.sav), Excel, Lotus, texto, etc. o desde base de datos como Access, dBase, FoxPro, etc. Si los datos son nuevos o se han modificado, es conveniente almacenarlos en disco a través del menú Archivo.

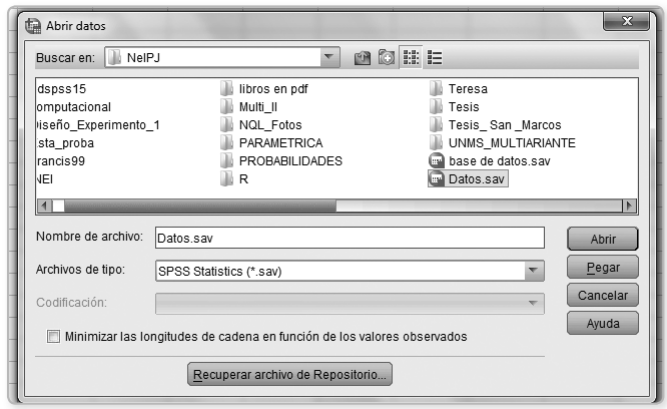
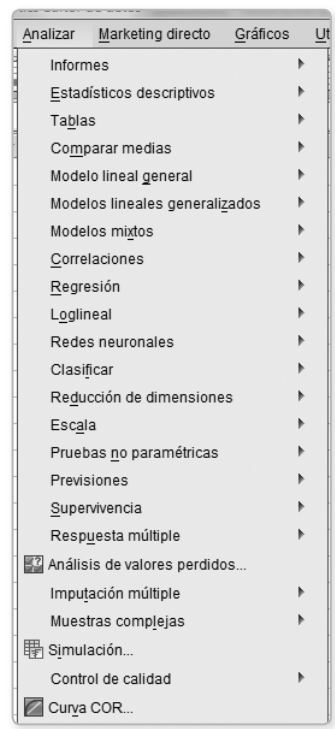



*Figura 1.1 Intentamos abrir el archivo Datos.sav*

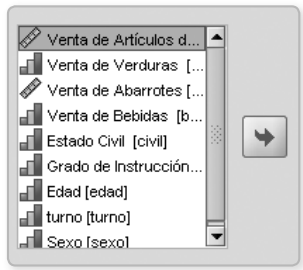
**B. Procedimiento**

Una vez ingresados los datos en el Editor de datos, se debe elegir un procedimiento en el menú Analizar, para poder obtener resultados. El análisis de datos dependerá del usuario, del tipo de la información y el conocimiento de la problemática en estudio. SPSS solo puede proveer las herramientas estadísticas que, correctamente utilizadas, permitan alcanzar el objetivo. El software le permite realizar estadísticas como informes, estadísticos, descriptivos, tablas, correlaciones, análisis de regresión, etc. Si se desea un gráfico descriptivo, se elige el menú Gráficos; puede ser normal o interactivo, en cuyo caso se podrá modificar posteriormente las variables usadas y los parámetros.



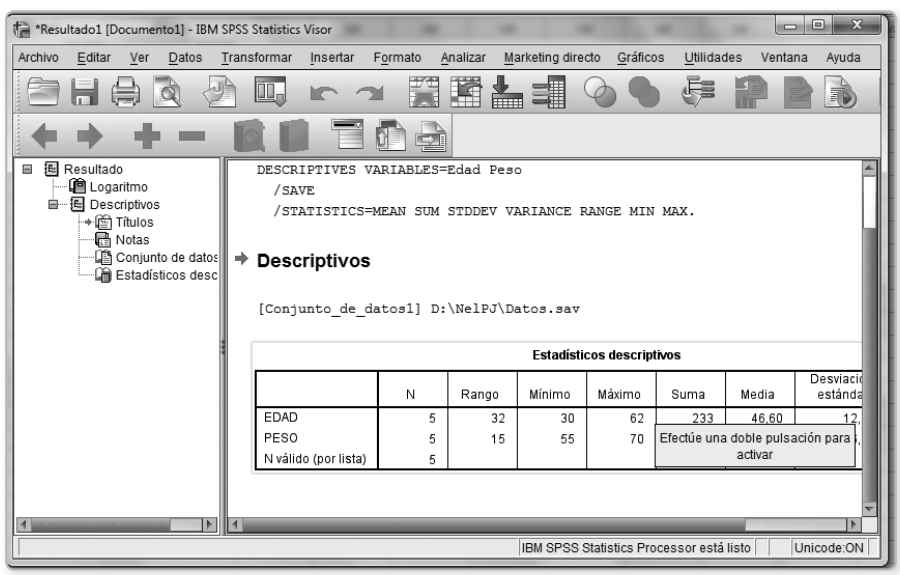
### C. Variables y opciones

Es fundamental seleccionar la variable o variables que se desea utilizar en el procedimiento de análisis o en el procedimiento de gráficar. El cuadro de selección de variables es similar en todos los procedimientos. Este cuadro muestra una lista de todas las variables y una o varias casillas de destino, donde debe introducirlas para que el sistema pueda calcular los estadísticos o valores que se precisen para la presentación de resultados.



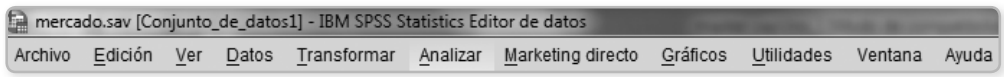

### D. Examen de resultados

El paso final es el examen de resultados en una ventana llamada Visor. Esta ventana muestra dos paneles: (1) el de la izquierda es el conjunto de los titulares de los elementos gráficos situados en el panel de la derecha. Su apariencia y funcionamiento es similar al del explorador de Windows. (2) El de la derecha muestra los elementos gráficos que pueden ser textos, tablas o gráficos. Los resultados o gráficos que se generan no se sustituyen, sino que se van acumulando en el visor donde se les puede dar formato para su posterior impresión o su almacenamiento en disco. En el caso de los gráficos interactivos, se puede modificar directamente, sin necesidad de volver a generar otra vez.



### 1.2.2. BARRA DE MENÚ

La barra de menú permite acceder a todas las funciones que implementa el SPSS. Cada ventana posee su propia barra de menús, pero varios de ellos son comunes a todas las ventanas. Los menús más importantes son los siguientes:




- **Archivo (File).** Permite abrir, cerrar, guardar e imprimir archivos. Es decir, agrupa todos los procedimientos involucrados en la creación de una nuevo archivo New, ya sea de datos SPSS (*.sav) o sintaxis (*.sbo) o leer archivos de bases de datos u hojas de cálculo, creados por otros programas de software.
- **Edición (Edit).** Para cortar, copiar y pegar elementos, además de acceder a las opciones generales. Permite editar y buscar datos, a partir de la ventana activa.
- **Ver (View).** Se puede hacer que se muestren u oculten diversos elementos del sistema. Además, permite personalizar la barra de herramientas, así como activar o desactivar el estado de esta.
- **Datos (Data).** Permite realizar cambios en los datos, como combinar ficheros, transponer variables y casos o crear grupos de casos para realizar cambios globales en el archivo de datos, tales como fusión de archivos, transposición y selección de variables y casos, segmentación de archivos, etc.
- **Transformar (Transform).** Permite modificar los datos de las variables, a partir de los ya existentes, en sí mismas o en otras nuevas; realiza cambios en variables seleccionadas en el archivo de datos y sirve para calcular nuevas variables basadas en los valores de las ya existentes.
- **Analizar (Analyze).** En este menú se encuentran los diferentes procedimientos de análisis estadístico para la elaboración de reportes y análisis de datos.
- **Marketing directo.** La opción Marketing directo ofrece un conjunto de herramientas diseñadas para mejorar el resultado de campañas de marketing directo, identificando y adquiriendo características que definen a diferentes grupos de consumidores, y dirigiéndose a grupos concretos para aumentar al máximo los índices de respuesta positivos.
- **Gráficos (Graphs).** Crea gráficos de barras, histogramas, etc., normales o interactivos, para variables cuantitativas y cualitativas.
- **Utilidades (Utilities).** Permite varias utilidades como definir conjuntos, ejecutar procesos, editar menús, cambiar la tipografía, migración de datos, crear un índice de comandos de SPSS.
- **Ventana (Window).** Permite disponer, seleccionar y controlar los atributos de las diferentes ventanas del SPSS. Además, permite gestión de ventanas abiertas y cambio de unas a otras.
- **Add ons (complementos).** Integra con todas las funciones, como el proceso de análisis, planificación, recopilación de datos, acceso a datos, gestión de datos y la preparación, análisis de datos, presentación de informes y el despliegue, etc.
- **Ayuda (Help).** Se trata de un acceso a la ayuda de SPSS, basada en temas. Brinda todos los temas expuestos en el formato de ayuda de Windows. Asimismo, presenta una ventana de ayuda estándar conteniendo información acerca de los procedimientos y funciones del SPSS.

La siguiente tabla resume las razones del por qué SPSS es su mejor elección para un análisis más preciso, comparado a una hoja de cálculo. Con la amplia gama de funcionalidades del SPSS y su fácil uso, podrá estar de acuerdo en que SPSS será el perfecto compañero de su hoja de cálculo.

| | |
|---|---|
| Necesita la mayor cantidad de información para tomar la mejor decisión. | SPSS va más allá que los gráficos estándares de una hoja de cálculo ,al poseer una mayor cantidad de cuadros de salida que le brindan mejores evidencias sobre sus datos, lo que le permite tomar decisiones con mayor confianza. |
| Existe más de una forma de analizar sus datos. | Al usar Pivot Tables de SPSS, finaliza los problemas de construcción manual de cuadros, recálculos y reformateo. Esto le otorga una manera fácil y libre de analizar sus datos. |
| El truco de un análisis efectivo radica en saber qué tan significativo es un resultado. | SPSS le da una amplia gama de pruebas estadísticas, que le permiten inferir si existen relaciones o diferencias significativas. |
| Es más fácil trabajar con palabras que con números. | SPSS brinda una presentación más comprensible, al utilizar las etiquetas definidas en cada variable en la presentación de cuadros y gráficos. |
| Es importante separar las manzanas de las naranjas. | SPSS le ahorra tiempo cuando necesita producir similares reportes y gráficos para diferentes subconjuntos de datos, con solo hacer clic. |
| Necesita resultados precisos, aún cuando existan datos perdidos. | Cuando ha perdido datos o desea cuantificar la significancia de las diferencias entre las no respuestas a una encuesta, SPSS automáticamente le da una mejor y más exacta información que una hoja de cálculo. |
| Percatarse cuando hay un problema con sus datos. | SPSS le ayuda a mostrar los errores de ingreso de datos o datos inusuales, que puede desear mantenerlos en su análisis o mostrarle un resultado previo antes de que escriba su informe final. |
| Importar datos en SPSS es fácil. | En SPSS es fácil importar datos de una hoja de cálculo, bases de datos y estructuras complejas. SPSS prácticamente elimina el riesgo de pérdidas accidentales de datos, mientras está haciendo algún análisis. |
| Todos tus datos son importantes. | SPSS puede trabajar fácilmente con 20 000 registros (filas) lo que una hoja de cálculo no, teniendo que fraccionar su archivo. |
| Usar la mejor herramienta para su trabajo le ahorrará tiempo e incrementará su productividad. | A diferencia de una hoja de cálculo que solo realiza operaciones matemáticas entre filas y columnas, SPSS le permite segmentar y agrupar grandes bases de datos de acuerdo con sus requerimientos de análisis. |
| Contestar sus preguntas debería ser fácil de encontrar y entender. | Una hoja de cálculo solo le ayuda indicándole los comandos que debe utilizar. SPSS le da una breve indicación de los términos estadísticos, reglas claves para interpretar determinados resultados. |
| Las herramientas estadísticas deberían mejorar de acuerdo con sus necesidades analíticas. | Una hoja de cálculo le limita a las herramientas estadísticas básicas. SPSS le permite ir añadiendo técnicas para análisis especializados, solo añadiendo módulos. |

### 1.2.3. BARRA DE HERRAMIENTAS

«[...] SPSS incorpora una o varias barras de herramientas según la ventana en la que nos encontremos. Estas barras contienen las funciones más usadas en el trabajo usual, para facilitar el acceso a ellas.»

Si detenemos el puntero del ratón unos segundos sobre una de las herramientas (o botones),nos presentará un pequeño cuadro con una breve descripción de la utilidad del mismo:



Las barras, además, son personalizables, pudiendo añadir o eliminar los botones que queramos así como moverlos; incluso permite cambiar de sitio la barra o borrarla completamente».

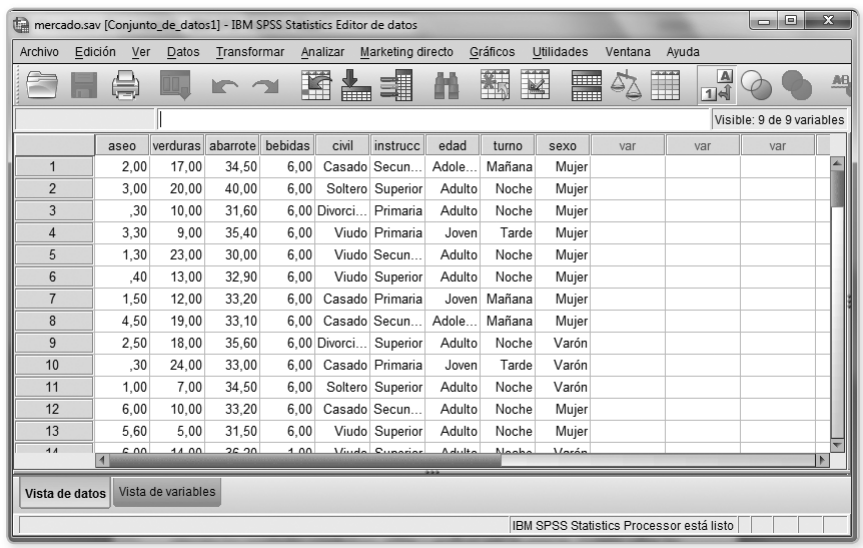
### 1.2.4. VENTANAS

Cuando se trabaja con SPSS, existen dos ventanas principales, que son el Editor de datos y el Visor; pero además, SPSS emplea otras ventanas secundarias, como la ventana de sintaxis y otras ventanas importantes.

A continuación, se presenta un resumen de las diferentes ventanas.

#### A. Editor de datos

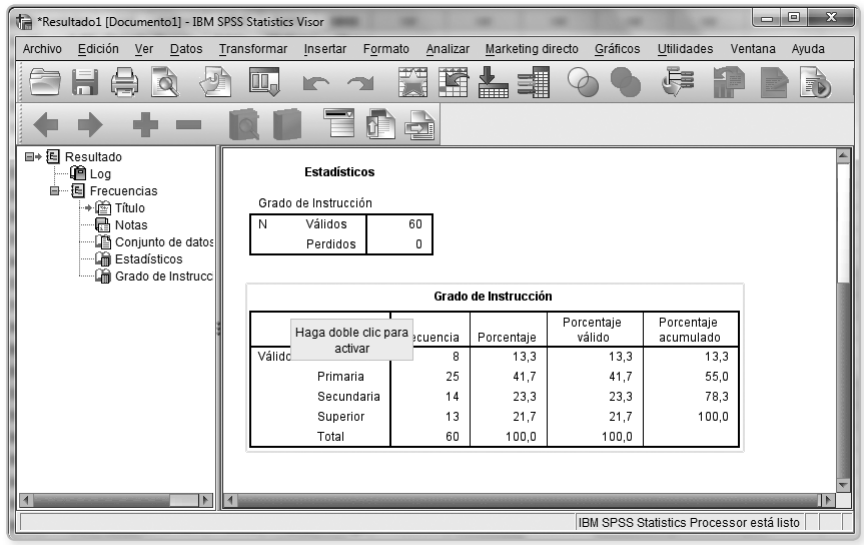
Muestra el contenido del fichero de datos. Permite crear nuevos archivos de datos o modificar los existentes. Esta ventana se abre automáticamente cuando se inicia una sesión de SPSS. No se puede tener más de un archivo de datos abierto al mismo tiempo; para ello, debemos abrir varias sesiones. Está compuesto de dos vistas: Vista de datos y Vista de variables.




| | aseo | verduras | abarrote | bebidas | civil | instrucc | edad | turno | sexo | var | var | var |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2,00 | 17,00 | 34,50 | 6,00 | Casado | Secun... | Adole... | Mañana | Mujer | | | |
| 2 | 3,00 | 20,00 | 40,00 | 6,00 | Soltero | Superior | Adulto | Noche | Mujer | | | |
| 3 | ,30 | 10,00 | 31,60 | 6,00 | Divorci... | Primaria | Adulto | Noche | Mujer | | | |
| 4 | 3,30 | 9,00 | 35,40 | 6,00 | Viudo | Primaria | Joven | Tarde | Mujer | | | |
| 5 | 1,30 | 23,00 | 30,00 | 6,00 | Viudo | Secun... | Adulto | Noche | Mujer | | | |
| 6 | ,40 | 13,00 | 32,90 | 6,00 | Viudo | Superior | Adulto | Noche | Mujer | | | |
| 7 | 1,50 | 12,00 | 33,20 | 6,00 | Casado | Primaria | Joven | Mañana | Mujer | | | |
| 8 | 4,50 | 19,00 | 33,10 | 6,00 | Casado | Secun... | Adole... | Mañana | Mujer | | | |
| 9 | 2,50 | 18,00 | 35,60 | 6,00 | Divorci... | Superior | Adulto | Noche | Varón | | | |
| 10 | ,30 | 24,00 | 33,00 | 6,00 | Casado | Primaria | Joven | Tarde | Varón | | | |
| 11 | 1,00 | 7,00 | 34,50 | 6,00 | Soltero | Superior | Adulto | Noche | Varón | | | |
| 12 | 6,00 | 10,00 | 33,20 | 6,00 | Casado | Secun... | Adulto | Noche | Mujer | | | |
| 13 | 5,60 | 5,00 | 31,50 | 6,00 | Viudo | Superior | Adulto | Noche | Mujer | | | |

### B. Visor

«Muestra los resultados estadísticos, tablas y gráficos que se generen. Permite editar los resultados, almacenarlos e imprimirlos. Esta ventana se abre automáticamente la primera vez que se ejecuta un procedimiento que genera resultados».




### C. Visor de borrador

«Esta ventana muestra los resultados como texto normal (en lugar de como tablas pivote interactivas)».

### D. Editor de tablas pivote

«Permite modificar los resultados mostrados en este tipo de tablas. Podemos editar el texto, intercambiar los datos de las filas y las columnas, añadir colores, crear tablas multidimensionales y ocultar y mostrar los resultados de manera selectiva.»

### E. Editor de gráficos

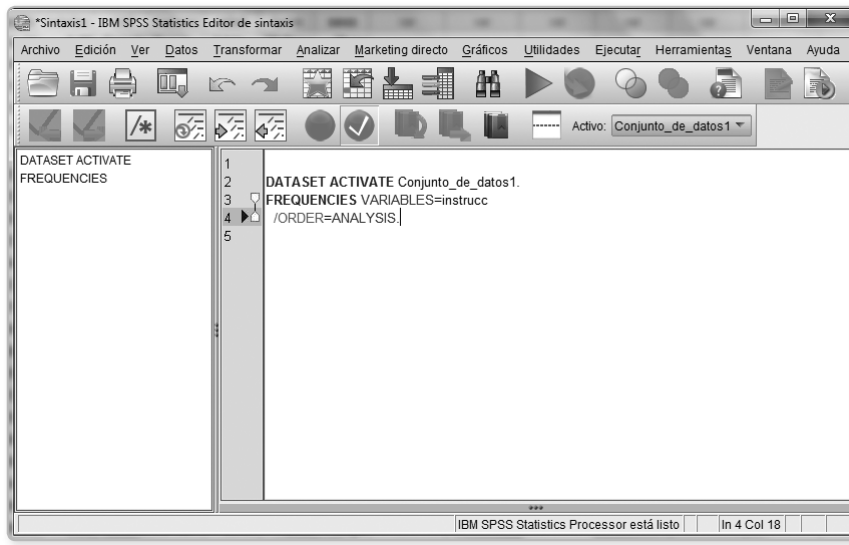
«Permite modificar los gráficos y diagramas. Es posible cambiar los colores, seleccionar diferentes tipos de fuentes y tamaños, intercambiar los ejes horizontal y vertical, rotar diagramas de dispersión 3-D e incluso cambiar el tipo de gráfico».

### F. Editor de resultados de texto

«Los resultados de texto que no aparecen en las tablas pivote pueden modificarse con este editor. Es posible variar el texto y cambiar las características de las fuentes (tipo, estilo, color y tamaño)».

### G. Editor de sintaxis

«Permite escribir conjuntos de comandos para que el sistema las ejecute (SPSS funciona internamente a base de comandos). Es posible pegar las selecciones de un cuadro de diálogo en una ventana de sintaxis, donde aparecerán en forma de comandos. Este editor permite utilizar las funciones especiales de SPSS que no se encuentran disponibles en los cuadros de diálogo. También puede guardar los comandos en un archivo para utilizarlos en sesiones posteriores».




### H. Editor de procesos

«El procesamiento y la automatización OLE permiten personalizar y automatizar muchas tareas en SPSS. Este editor permite crear y modificar procesos básicos.»

CAPÍTULO

2

# MANEJO DE DATOS

En este capítulo se abordará la parte operativa más importante al momento de realizar un trabajo, es decir, aprender a manejar la ventana Editor de datos. Se presentan dos situaciones en la edición de datos: primero, si la información provienen de una encuesta, cuestionario u otro tipo de formato, entonces los datos se introducirán directamente a la ventana Editor de datos (tabla que está constituida por un conjunto de datos organizados en una estructura matricial, donde las filas constituyen los casos y las columnas las variables). La segunda forma de contar con información es extrayendo datos de un archivo de disco (USB, CD, DVD, disco magnético flexible, disco duro, PD, base de datos, etc.).

Existen dos maneras de obtener información y trabajar en el editor de datos: la primera es la *introducción directa de los datos* en las celdas que muestra el Editor de datos. La segunda es obteniendo datos *extrayendo un archivo de disco*.

## 2.1. INTRODUCCIÓN DIRECTA DE DATOS

Tanto los cuestionarios como los demás formatos de recolección de información constituyen una forma estructurada, sistemática y organizada para registrar los datos. Construir un formato de recolección de información o cuestionario obligará al investigador a pensar en términos de sus requerimientos de información, al planeamiento del trabajo de campo o al diseño de su experimento, según sea el caso y las técnicas estadísticas para el análisis de los resultados.

La introducción directa de datos sigue tres pasos importantes: realizar el cuestionario, codificar el cuestionario y llenar la base de datos.

### Ejemplo

Se tiene la siguiente encuesta (cuestionario):

**Encuesta sobre compras en el supermercado**

Se quiere realizar un análisis del comportamiento de los consumidores que realizaron sus compras semanales en un supermercado, realizando el siguiente cuestionario. El cuestionario consta de nueve preguntas, las primeras cinco (1 al 5) son variables de cualidad y las cuatro últimas (6-9) son de cantidad.

Marque con una "x" la alternativa correcta o escriba en las líneas punteadas.

1. Edad
   a) Adolescente ☐ b) Joven ☐ c) Adulto ☐
2. Sexo
   a) Varón ☐ b) Mujer ☐
3. Grado de instrucción
   a) Primaria ☐ b) Secundaria ☐ c) Superior ☐ d) No tiene ☐
4. Estado civil
   a) Soltero ☐ b) Casado ☐ c) Divorciado ☐ d) Viudo ☐
5. Horario de compra
   a) Mañana ☐ b) Tarde ☐ c) Noche ☐
6. Compras semanales en artículos de aseo personal (nuevo soles) .....
7. Compras semanales en verduras (nuevo soles) .....
8. Compras semanales en abarrotes (nuevo soles) .....
9. Compras semanales en bebidas (nuevo soles) .....

### 2.1.1 REALIZAR LA ENCUESTA

Primero se debe definir cuántas encuestas se van a realizar; es decir, hallar el tamaño de muestra. En el ejemplo se tomó una muestra de 30 casos (individuos, sujetos, objetos, etc.). En segundo lugar, se recopilan los datos correspondientes a las 30 encuestas; es decir, se realiza la encuesta a las 30 personas. Por otro lado, puesto que los datos constituyen una muestra aleatoria de los clientes del supermercado, la interpretación de los análisis realizados se referirá tan solo a esta muestra de sujetos.

### 2.1.2 CODIFICAR LA ENCUESTA

Las variables del cuestionario (preguntas) deben ser codificadas utilizando códigos convenientes. Se recomiendan números enteros positivos, incluido el cero. Así, la primera variable Edad se puede codificar de la siguiente manera: a la alternativa Adolescente se le asigna el código "0"; a Joven, el código "1"; y Adulto, código "2". De la misma forma, se puede seguir asignando códigos a cada atributo de las variables (preguntas). Esta codificación solo es posible si la pregunta es de cualidad. En consecuencia, solo se podrá codificar hasta la pregunta cinco, debido que a partir de la pregunta seis las preguntas son de cantidad y no necesitan ser codificados, tal como se muestra a continuación:

1: Edad (0: Adolescente; 1: Joven; 2: Adulto)

2: Sexo (0: Mujer; 1: Varón)

3: Grado de instrucción (1: No tiene; 2: Primaria; 3: Secundaria; 4: Superior)

4: Estado civil (1: Soltero; 2: Casado; 3: Divorciado; 4: Viudo)

5: Turno (0: Mañana; 1: Tarde; 2: Noche)

6: Venta de artículos de aseo personal (nuevos soles)

7: Venta de verduras (nuevos soles)

8: Venta de abarrotes (nuevos soles)

9: Venta de bebidas (nuevo soles)

La tabla representa las encuestas codificadas, donde las columnas indican las variables y las filas los casos.

| Ni | Preg. 1 | Preg. 2 | Preg. 3 | Preg. 4 | Preg. 5 | Preg. 6 | Preg. 7 | Preg. 8 | Preg. 9 |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 0 | 0 | 3 | 2 | 0 | 2,0 | 17,0 | 34,5 | 6,0 |
| **2** | 2 | 0 | 4 | 1 | 2 | 3,0 | 20,0 | 40,0 | 6,0 |
| **3** | 2 | 0 | 2 | 3 | 2 | 0,3 | 10,0 | 31,6 | 6,0 |
| **4** | 1 | 0 | 2 | 4 | 1 | 3,3 | 9,0 | 35,4 | 6,0 |
| **5** | 2 | 0 | 3 | 4 | 2 | 1,3 | 23,0 | 30,0 | 6,0 |
| **6** | 2 | 0 | 4 | 4 | 2 | 0,4 | 13,0 | 32,9 | 6,0 |
| **7** | 1 | 0 | 2 | 2 | 0 | 1,5 | 12,0 | 33,2 | 6,0 |
| **8** | 0 | 0 | 3 | 2 | 0 | 4,5 | 19,0 | 33,1 | 6,0 |
| **9** | 2 | 1 | 4 | 3 | 2 | 2,5 | 18,0 | 35,6 | 6,0 |
| **10** | 1 | 1 | 2 | 2 | 1 | 0,3 | 24,0 | 33,0 | 6,0 |
| **11** | 2 | 1 | 4 | 1 | 2 | 1,0 | 7,0 | 34,5 | 6,0 |
| **12** | 2 | 0 | 3 | 2 | 2 | 6,0 | 10,0 | 33,2 | 6,0 |
| **13** | 2 | 0 | 4 | 4 | 2 | 5,6 | 5,0 | 31,5 | 6,0 |
| **14** | 2 | 1 | 4 | 4 | 2 | 6,0 | 14,0 | 36,2 | 1,0 |
| **15** | 1 | 0 | 1 | 1 | 1 | 1,2 | 15,0 | 36,8 | 1,0 |
| **16** | 2 | 0 | 1 | 2 | 2 | 0,2 | 12,0 | 35,4 | 2,0 |
| **17** | 2 | 0 | 1 | 3 | 2 | 6,0 | 14,0 | 33,2 | 2,0 |
| **18** | 2 | 0 | 2 | 2 | 2 | 5,5 | 6,0 | 35,9 | 2,0 |
| **19** | 2 | 0 | 4 | 4 | 2 | 6,5 | 9,0 | 36,5 | 3,0 |
| **20** | 2 | 0 | 1 | 4 | 2 | 0,2 | 13,0 | 38,2 | 3,0 |
| **21** | 2 | 0 | 3 | 1 | 2 | 2,3 | 6,0 | 34,5 | 3,0 |
| **22** | 2 | 0 | 2 | 2 | 2 | 0,2 | 7,0 | 36,2 | 3,0 |
| **23** | 1 | 0 | 1 | 1 | 1 | 2,3 | 10,0 | 36,8 | 3,0 |
| **24** | 2 | 0 | 2 | 4 | 2 | 1,5 | 13,0 | 30,1 | 3,0 |
| **25** | 2 | 0 | 3 | 4 | 2 | 5,0 | 7,0 | 36,0 | 3,0 |
| **26** | 2 | 0 | 2 | 2 | 2 | 4,5 | 18,0 | 35,4 | 3,0 |
| **27** | 0 | 0 | 3 | 2 | 0 | 0,3 | 24,0 | 31,5 | 3,0 |
| **28** | 2 | 0 | 2 | 3 | 2 | 5,5 | 7,0 | 36,2 | 4,0 |
| **29** | 2 | 0 | 2 | 2 | 2 | 0,5 | 10,0 | 33,9 | 4,0 |
| **30** | 2 | 0 | 2 | 2 | 2 | 0,2 | 4,0 | 34,9 | 4,0 |

**NOTA**

*En definitiva, para trabajar con el SPSS no es necesario codificar todas las encuestas, es suficiente con una y esta servirá de modelo para trabajar en el editor de datos al momento de ingresar la información.*

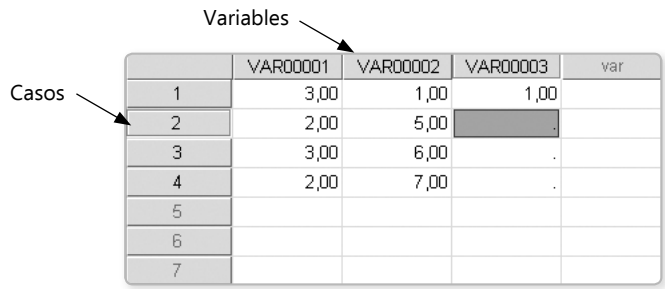
### 2.1.3 LLENAR LA BASE DE DATOS (EDITOR DE DATOS)

Una vez codificada la encuesta se debe ingresar la información al editor de datos. Cuando se recolectan los datos de n casos (sujetos) en j variables (preguntas), debe introducirlos en la matriz de datos de SPSS. Para ello, sitúese en la ventana Editor de datos; esta ventana contiene Vista de datos (*Data View*) y Vista de variables (*Variable View*). A continuación, se describen las pestañas Vista de datos y Vista de variables.

En este caso, el valor de n = 30 encuestados y j= 9 preguntas.

#### A. Vista de datos (Data View): Ingreso de datos

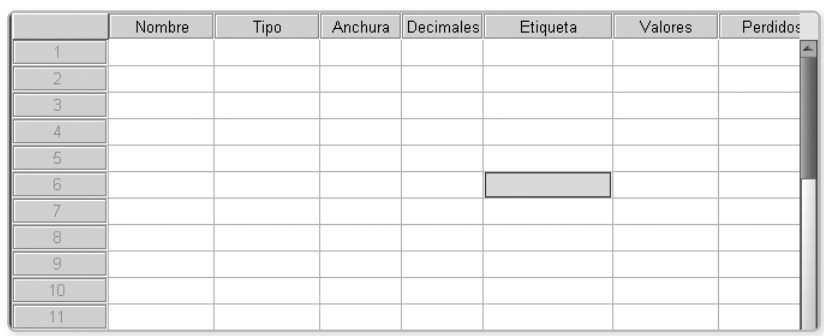
Vista de datos *(Data View)* presenta una *matriz de "n" casos (sujetos) x "j"* variables Cada columna representa una variable única y cada fila un caso individual de cada variable. Para introducir datos, solo se necesita pinchar sobre una casilla y teclear el dato. Si donde se teclea el dato la variable no está definida, SPSS asigna automáticamente un nombre de variable a la columna y define por defecto sus parámetros.




Si bien es cierto se puede empezar a introducir directamente los datos en la matriz de datos (Vista de datos), es recomendable primero definir las variables en Vista de variables.

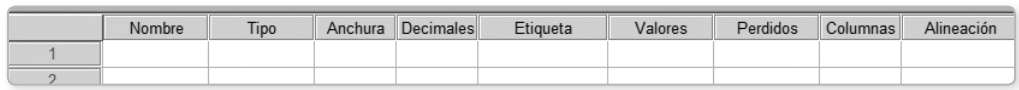
#### B. Vista de variables (Variable View): Definir variables

Si se pulsa en la pestaña Vista de variables, se accede a una matriz similar a la de datos; pero, en este caso, contiene las variables definidas del modo siguiente: cada fila representa una variable (1, 2, 3...) y cada columna un parámetro o especificación de esa variable (Name, Type, etc.). Estos se asignan por defecto; sin embargo, puede modificarlos a su conveniencia. La lista de parámetros o especificación de variables es la siguiente:

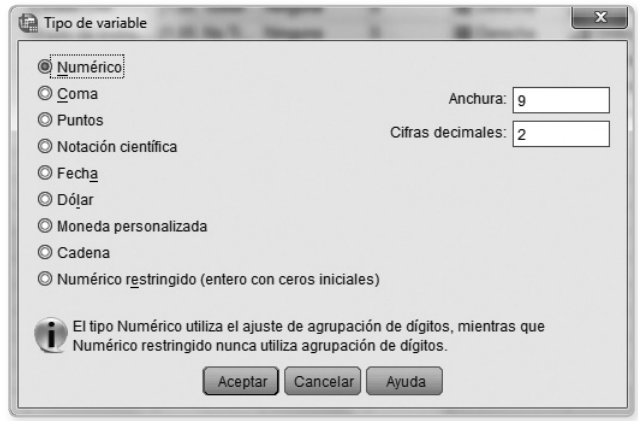


A continuación, se describe cada uno de los parámetros:

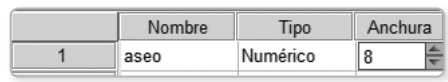
- **Nombre (*Name*).** Introduzca el nombre de la variable, que pueden ser letras, números o el símbolo de subrayado (_). El primer carácter únicamente puede ser una letra y no se hacen distinciones entre mayúsculas y minúsculas. Por ejemplo, en la fila "1", escriba Edad para el parámetro Name (Nombre).



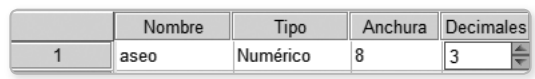

- **Tipo (*Type*)**. Si se ubica en el parámetro Type (Tipo), a través del botón ··· que aparece al lado de tipo (Numeric), se accede a un panel donde se puede seleccionar el tipo de variable a utilizar, que puede ser numérica, coma, fecha, moneda, cadena, etc.



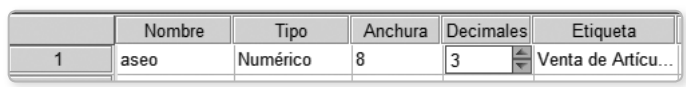

- **Anchura (*Width*).** El parámetro Anchura (Width) especifica el número de caracteres o dígitos de la variable. El botón , que aparece al lado izquierdo permite aumentar o disminuir el número de dígitos o caracteres.



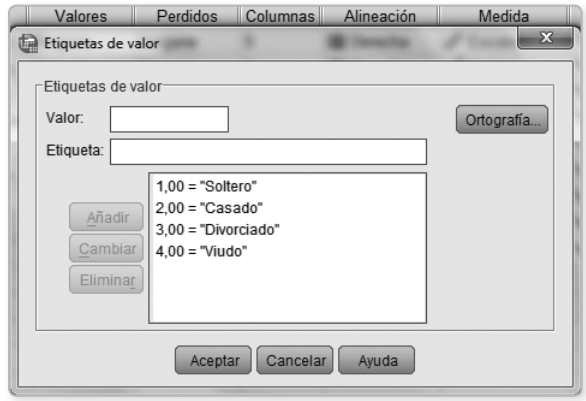

- **Decimales (*Decimals*).** El parámetro Decimales (Decimals) indica el número de decimales o si la variable es de tipo contable (numérico, moneda, etc.). El botón que aparece al lado izquierdo permite aumentar o disminuir el número de decimales




- **Etiqueta (*Label*).** Se usan etiquetas (*Labels*) para describir la variable de una forma más clara. Las etiquetas se utilizan en la presentación de resultados.



- **Valores (*Values*).** El parámetro Valores (*Values*) permite asignar etiquetas a determinados valores de las variables. Por ejemplo, la variable numérica Edad contiene los siguientes valores: 0 si es adolescente, 1 si es joven y 2 si es adulto; entonces, para el ***valor (value) 0***, su etiqueta de valor (*value label*) será Adolescente, luego haga clic en Add. Para el ***valor 1***, su etiqueta de valor será Joven y para ***valor 2***, su etiqueta de valor será Adulto, respectivamente. De esta forma, cuando se trabaja con la variable Edad aparecerá Adolescente, Joven y Adulto, en lugar de 0, 1 y 2.



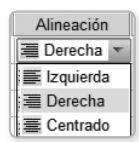

- **Perdidos (*Missing*).** El parámetro Perdidos (*Missing*) permite definir los valores que se tratarán como perdidos. Se debe distinguir entre los valores declarados por el sistema (ausencia de datos) y los valores declarados como perdidos por uno mismo.

| | Nombre | Tipo | Anchura | Decimales | Etiqueta | Valores | Perdidos |
|---|---|---|---|---|---|---|---|
| 1 | aseo | Numérico | 8 | 3 | Venta de Artícu... | Ninguna | Ninguna |

- **Columnas (*Columns*).** El parámetro Columnas (*Columns*) indica lel ancho de la columna en la Vista de datos. El botón que aparece al lado izquierdo permite aumentar o disminuir el ancho.

| | Nombre | Tipo | Anchura | Decimales | Etiqueta | Valores | Perdidos | Columnas |
|---|---|---|---|---|---|---|---|---|
| 1 | aseo | Numérico | 8 | 3 | Venta de Artícu... | Ninguna | Ninguna | 5 |

- **Alineación (*Aling*).** El parámetro Alineación indica la posición del dato en la casilla de Vista de datos, la misma que puede ser alineación Izquierda (*Left*), alineación Derecha (*Righ*) o Centrado (*Center*).




- **Medida (*Measure*).** Este parámetro permite especificar el nivel de medida, que puede ser ***Escala*** (datos numéricos de una escala de intervalo o de razón), ***Ordinal*** (representan categorías ordenadas) y ***Nominal*** (representan categorías sin orden alguno). Esta especificación solo se usa en procedimientos gráficos.



En general, es importante saber distinguir entre una variable categórica y una de escala. Una **variable categórica** es aquella que tiene un número limitado de valores o categorías distintas. Todas las variables de cadena y las variables numéricas con etiquetas de valor definidas, o las variables numéricas definidas como nominales u ordinales se tratan como categóricas. El resto son tratadas como de escala.

## Ejemplo:

En la encuesta realizada en el supermercado, defina las nueve variables del cuestionario.

## Solución:

Las variables a definir serán nueve: edad, sexo, grado de instrucción, etc. En cada fila se escribirá el **nombre** de la variable (teniendo en cuenta que no puede tener muchos caracteres de largo), además de otras propiedades como el **tipo** (numérico, cadena, etc.), la **anchura**, el número de **decimales**, la **etiqueta**, el "valor" y la "etiqueta de valor" de los **valores** que admite los valores **perdidos**, la **alineación** del texto, etc. Solo posicione el mouse o el cursor en cada casilla y escriba la información deseada. En el ejercicio popuesto, la respuesta se muestra en el gráfico siguiente:

| | Nombre | Tipo | Anchura | Decimales | Etiqueta | Valores | Perdidos | Columnas | Alineación | Medida |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | edad | Numérico | 9 | 2 | Edad | {,00, Adoles... | Ninguna | 5 | Izquierda | Ordinal |
| 2 | sexo | Numérico | 9 | 2 | Sexo | {,00, Mujer}... | Ninguna | 5 | Izquierda | Ordinal |
| 3 | instrucc | Numérico | 9 | 2 | Grado de Instru... | {1,00, No Ti... | Ninguna | 7 | Izquierda | Ordinal |
| 4 | civil | Numérico | 9 | 2 | Estado Civil | {1,00, Solter... | Ninguna | 5 | Izquierda | Ordinal |
| 5 | turno | Numérico | 9 | 2 | turno | {,00, Mañan... | Ninguna | 5 | Izquierda | Ordinal |
| 6 | aseo | Numérico | 8 | 3 | Venta de Artícu... | Ninguna | Ninguna | 5 | Izquierda | Escala |
| 7 | verduras | Numérico | 9 | 2 | Venta de Verdu... | Ninguna | Ninguna | 5 | Izquierda | Ordinal |
| 8 | abarrote | Numérico | 9 | 2 | Venta de Abarr... | Ninguna | Ninguna | 5 | Izquierda | Escala |
| 9 | bebidas | Numérico | 9 | 2 | Venta de Bebidas | Ninguna | Ninguna | 5 | Izquierda | Ordinal |

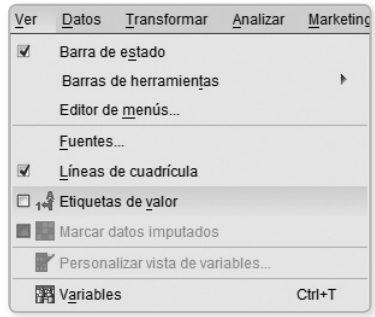
### C. Ver datos (Vista de datos)

Definidas las variables, se pasa a Vista de datos para introducir los datos en la matriz. Para desplazarse por la matriz de datos use las flechas del cursor, así como la tecla Intro para pasar al siguiente caso de la misma variable, y la tecla Tab para pasar a la siguiente variable.

| | edad | sexo | instrucc | civil | turno | aseo | verduras | abarrote | bebidas |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 3 | 2 | 0 | 2,00 | 17,00 | 34,50 | 6,00 |
| 2 | 2 | 0 | 4 | 1 | 2 | 3,00 | 20,00 | 40,00 | 6,00 |
| 3 | 2 | 0 | 2 | 3 | 2 | ,30 | 10,00 | 31,60 | 6,00 |
| 4 | 1 | 0 | 2 | 4 | 1 | 3,30 | 9,00 | 35,40 | 6,00 |
| 5 | 2 | 0 | 3 | 4 | 2 | 1,30 | 23,00 | 30,00 | 6,00 |
| 6 | 2 | 0 | 4 | 4 | 2 | ,40 | 13,00 | 32,90 | 6,00 |
| 7 | 1 | 0 | 2 | 2 | 0 | 1,50 | 12,00 | 33,20 | 6,00 |
| 8 | 0 | 0 | 3 | 2 | 0 | 4,50 | 19,00 | 33,10 | 6,00 |
| 9 | 2 | 1 | 4 | 3 | 2 | 2,50 | 18,00 | 35,60 | 6,00 |
| 10 | 1 | 1 | 2 | 2 | 1 | ,30 | 24,00 | 33,00 | 6,00 |
| 11 | 2 | 1 | 4 | 1 | 2 | 1,00 | 7,00 | 34,50 | 6,00 |
| 12 | 2 | 0 | 3 | 2 | 2 | 6,00 | 10,00 | 33,20 | 6,00 |
| 13 | 2 | 0 | 4 | 4 | 2 | 5,60 | 5,00 | 31,50 | 6,00 |
| 14 | 2 | 1 | 4 | 4 | 2 | 6,00 | 14,00 | 36,20 | 1,00 |
| 15 | 1 | 0 | 1 | 1 | 1 | 1,20 | 15,00 | 36,80 | 1,00 |
| 16 | 2 | 0 | 1 | 2 | 2 | ,20 | 12,00 | 35,40 | 2,00 |
| 17 | 2 | 0 | 1 | 3 | 2 | 6,00 | 14,00 | 33,20 | 2,00 |
| 18 | 2 | 0 | 2 | 2 | 2 | 5,50 | 6,00 | 35,90 | 2,00 |

*Figura 2.1 El gráfico muestra los datos en forma de valor original.*

Si necesita los datos definidos con etiquetas, diríjase a la opción del menú Ver (*View*) > Etiquetas de valor (*Value Labels*).



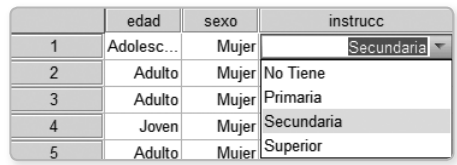

En todo caso, también con el botón Value Labels (Etiquetas de valor) se pueden mostrar las etiquetas en lugar del valor original.



Luego, los valores de datos definidos con etiquetas le permiten introducirlos como tales o con el nombre de la etiqueta, como se puede observar en la siguiente matriz de datos:

| | edad | sexo | instrucc | civil | turno | aseo | verduras | abarrote | bebidas |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Adole... | Mujer | Secundaria | Casado | Mañana | 2,00 | 17,00 | 34,50 | 6,00 |
| 2 | Adulto | Mujer | Superior | Soltero | Noche | 3,00 | 20,00 | 40,00 | 6,00 |
| 3 | Adulto | Mujer | Primaria | Divorci... | Noche | ,30 | 10,00 | 31,60 | 6,00 |
| 4 | Joven | Mujer | Primaria | Viudo | Tarde | 3,30 | 9,00 | 35,40 | 6,00 |
| 5 | Adulto | Mujer | Secundaria | Viudo | Noche | 1,30 | 23,00 | 30,00 | 6,00 |
| 6 | Adulto | Mujer | Superior | Viudo | Noche | ,40 | 13,00 | 32,90 | 6,00 |
| 7 | Joven | Mujer | Primaria | Casado | Mañana | 1,50 | 12,00 | 33,20 | 6,00 |
| 8 | Adole... | Mujer | Secundaria | Casado | Mañana | 4,50 | 19,00 | 33,10 | 6,00 |
| 9 | Adulto | Varón | Superior | Divorci... | Noche | 2,50 | 18,00 | 35,60 | 6,00 |
| 10 | Joven | Varón | Primaria | Casado | Tarde | ,30 | 24,00 | 33,00 | 6,00 |
| 11 | Adulto | Varón | Superior | Soltero | Noche | 1,00 | 7,00 | 34,50 | 6,00 |
| 12 | Adulto | Mujer | Secundaria | Casado | Noche | 6,00 | 10,00 | 33,20 | 6,00 |
| 13 | Adulto | Mujer | Superior | Viudo | Noche | 5,60 | 5,00 | 31,50 | 6,00 |
| 14 | Adulto | Varón | Superior | Viudo | Noche | 6,00 | 14,00 | 36,20 | 1,00 |
| 15 | Joven | Mujer | No Tiene | Soltero | Tarde | 1,20 | 15,00 | 36,80 | 1,00 |
| 16 | Adulto | Mujer | No Tiene | Casado | Noche | ,20 | 12,00 | 35,40 | 2,00 |

Al seleccionar una casilla, una lista desplegable aparece con todas las opciones que se tienen para la variable. Para modificar un dato es suficiente con pinchar su casilla y editarlo, bien en la misma casilla, bien en el Editor de casillas (situado entre la barra de herramientas y el Editor de datos). Se debe tener en cuenta que en el Editor de casillas no se puede introducir alguna fórmula u operación.




Es importante guardar la información introducida en el Editor de datos en una unidad de disco, a través del menú Archivo de la barra de herramientas, o seleccionando el menú Archivo; luego, se elige Guardar y aparece el correspondiente cuadro de diálogo.

**D. Uso del Editor de datos y variables**

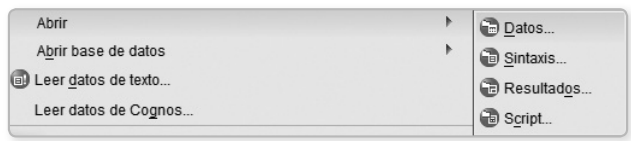
El Editor de datos permite realizar múltiples operaciones con la matriz de datos. Básicamente, con las operaciones cortar, copiar y pegar del menú Edición se puede realizar cualquier desplazamiento o copia de los datos de manera libre.

Puede copiar un dato y pegarlo en varias celdas, o copiar uno o varios casos completos en otra u otras filas, copiar grupos de casos de una o varias variables, o variables completas, sustituyendo a otras y creando otras nuevas. En resumen, casi cualquier operación en la matriz de datos es posible.
Los datos que no le proporcione directamente el sistema los completará con los valores definidos por defecto. En el caso de la ventana de variables, el funcionamiento es el mismo, con la restricción de que no se pueden modificar las especificaciones definidas de las variables. Asimismo, se pueden insertar casos y variables mediante el menú Datos, mediante los botones, desplazándose a un caso determinado con el comando Ir a caso o con una variable.

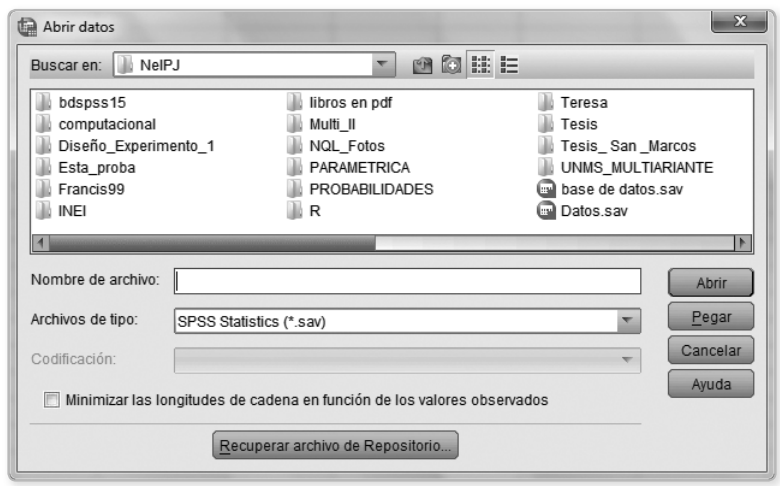
## 2.2. EXTRAER UN ARCHIVO DE DISCO

Otra forma de obtener datos es extrayendo un archivo de disco llamado «soporte electrónico de datos».Se trata de todo documento o elemento que conteniendo información es "leído" o interpretado por la unidad periférica de un ordenador (USB, CD, disco magnético flexible, disco duro, PD, etc.). Así, la información recogida en un cuestionario o formato de recolección debe ser registrada en un soporte, para que la información pueda ser procesada por un ordenador.

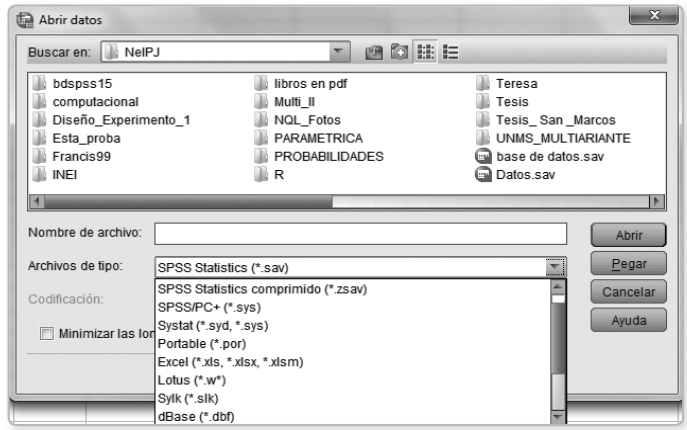
El procedimiento consiste en ir a la opción del menú Archivo (*File*) > Abrir (*Open*) > Datos (*Data*).




Luego, ubique la unidad de disco o carpeta donde se encuentra el archivo que se desea extraer.



Este archivo puede ser de formato (*.sav), Excel (*.xls), Lotus, Texto, etc. o de base de datos como Access, dBase, FoxPro, etc.; inclusive base de datos corporativos o descargados de la Web.



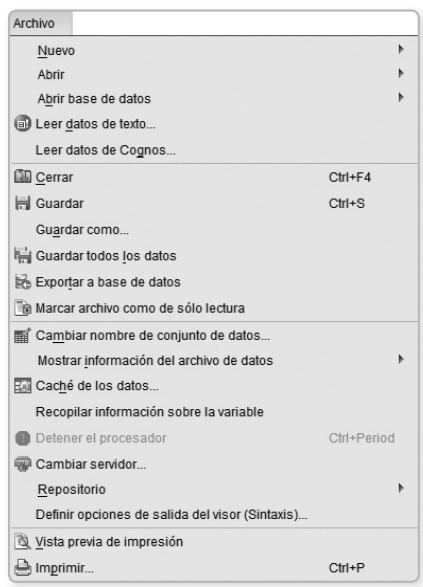

## 2.2.1. UN PASEO POR EL MENÚ

La barra de menús consta de una serie de opciones (Archivo, Edición, Ver, Datos, etc.). Si se selecciona cada una de ellas, se muestra un menú desplegable donde se ofrecen otros submenús, cada uno de los cuales tiene a su vez un cuadro de diálogo. Entre los menús, se encuentran los que son comunes a todos los programas de tipo Windows: Archivar, Editar, Ver, Ventana, etc. Los que ofrecen mayores diferencias son los menús Datos, Transformar y Analizar, los cuales permiten realizar operaciones con los datos.




### A. Menú Archivo (File)

Dentro del menú Archivo *(File)*, se puede acceder a multitud de operaciones, como Nuevo, Abrir, Guardar, etc.



- **Nuevo (*New*).** La primera opción es crear un nuevo archivo, ya sea de datos (*.sav), sintaxis (*.sps), resultados (*.spo), resultados de borrador (*.rtf) o procesos (*.sbs). En una misma sesión no se puede tener más de un archivo de datos abierto simultáneamente.
- **Abrir (*Open*).** La segunda opción es abrir un archivo existente en un disco de cualquiera de los tipos mencionados; pero en este caso, no solo se podrá leer datos propios de SPSS; sino que se puede acceder a otros tipos de archivo como los pertenecientes a Excel, Lotus, dBase, etc.
- **Abrir base de datos (*Open Database*).** También se puede obtener un fichero de datos a partir de una base de datos externa. Para ello, usted dispone de un asistente para realizar consultas sobre la base de datos y obtener la matriz de datos que precise.
- **Datos de ficheros de texto (*Read Text Data*).** La última opción para obtener datos es la de ficheros de texto. Para importar con éxito un fichero de texto, siga una serie de pasos guiados por un asistente.
- **Guardar (*Save*).** La siguiente opción sirve para guardar un archivo nuevo o seguir almacenando información ya existente con el mismo formato.
- **Guardar como (*Save As...*).** Esta opción le permite almacenar sus datos en disco, según el formato propio o de Excel, dBase, Lotus, texto, etc. También se utiliza para cambiar el nombre de un archivo.
- **Mostrar información de datos (*Display Data Info...*).** Presenta a través del Visor la información de cualquier archivo de disco. Consiste en obtener información relativa al archivo *.sav y a las variables que lo componen, antes de abrir el archivo *.sav.
- **Aplicar diccionario de datos.** La información del archivo de datos de disco que le solicita, sean etiquetas, valores perdidos y formatos, se aplica al archivo de trabajo. Esta información se aplica basándose en los nombres de las variables coincidentes, de modo que el resto de variables permanecen intactas.
- **Hacer caché de datos.** Crea una copia temporal de la tabla de datos en disco. Esto puede ser interesante para acelerar las lecturas de los datos, siempre que se tenga espacio suficiente en el disco.
- **Imprimir.** Permite imprimir cuadros, gráficos o textos; es decir, se puede imprimir desde la ventana de resultados, la ventana de sintaxis etc.
- **Presentación preliminar.** Se muestra una vista preliminar de cómo quedarán los datos, gráficos, tablas y textos una vez impresos.
- **Cambiar servidor.** El análisis en modo distribuido permite utilizar un equipo remoto, normalmente de mayor potencia, para realizar trabajos de gran esfuerzo computacional. Para ello, se requiere de una versión local (o cliente) en su equipo y de una versión servidor en el equipo remoto. Pulse en la opción Cambiar servidor, seleccione el servidor remoto al que desea conectarse e introduzca el identificador (ID), la contraseña y el dominio (si es necesario). Una vez conectados, se abrirá un archivo remoto a través del servidor.

  Si desea un archivo de su propio equipo, debe llegar hasta él a través del servidor (ruta de acceso UNC, convención de denominación universal). El resto del análisis distribuido lo hará el sistema por sí solo, de forma transparente al usuario.

### B. Menú Edición (Edit)

El menú Edición presenta opciones de mucha ayuda en la manipulación de los datos en estudio.




- **Deshacer (Ctrl+Z) y rehacer (Ctrl+R)**. Que permanecerá desactivada hasta que deshaga algún error. No debe confiarse siempre a la opción de deshacer; ya que no siempre se permite ejecutarla, sobre todo cuando se actúa directamente sobre el fichero de disco.
- **Cortar, copiar, pegar, eliminar, etc.** Estas opciones son las usuales de cualquier programa basado en el entorno Windows.
- **Buscar.** Esta opción permite encontrar cualquier conjunto de caracteres numéricos o alfanuméricos en la matriz de datos.
- **Opciones.** Presenta un panel con las especificaciones generales. Permite cambiar los directorios usados, los sistemas de medida, las diferentes configuraciones y formatos de visores, gráficos, tablas, datos, procesos, etc.

### C. Menú Ver (View)

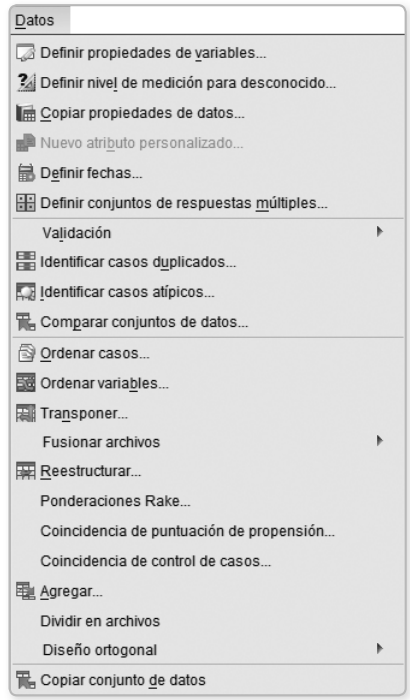
A través del menú Ver se accede al formato de visualización de los datos, como la fuente empleada o la visibilidad de la cuadrícula o de la barra de estado. Esta última (situada en la parte inferior de la ventana) muestra información relativa al sistema y a los datos.




Además, se puede configurar la barra de herramientas, crear nuevas y personalizarlas libremente, o conmutar entre la Vista de datos y la Vista de variables.

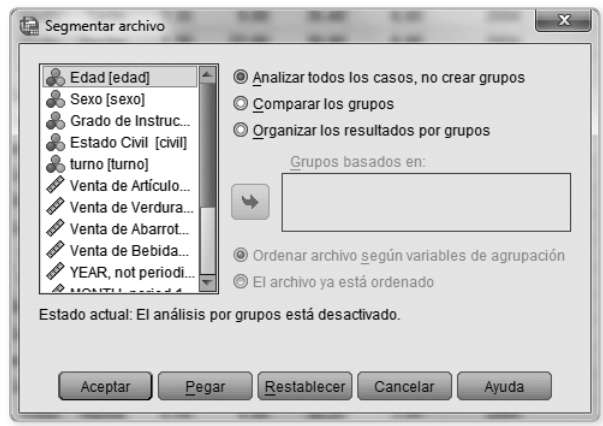
### D. Menú Datos (Data)

Las funciones más importantes del menú datos son las siguientes:




- **Definir fechas.** Genera variables de fecha que se utilizan para establecer la periodicidad de una serie temporal (cada caso corresponde a un momento de tiempo distinto y deben espaciarse de forma uniforme), y para etiquetar los resultados de los análisis de series temporales. Genera una anotación en el Visor, además de las ya comentadas opciones de inserción y desplazamiento (manipulación de datos).
- **Ordenar casos**. Ordenar los casos por una o más variables de forma ascendente o descendente.
- **Transponer.** Crea un archivo de datos nuevo en el que se transponen las filas y las columnas del archivo de datos original, de manera que los casos (las filas) se convierten en variables y las variables (las columnas) en casos. Los nombres de las nuevas variables se crean automáticamente si no se les asigna una variable determinada en la que todos sus casos sean diferentes. Una vez realizada la transposición, aparecerá una variable llamada *case_lbl,* cuyos casos serán los nombres de las antiguas variables.
- **Fundir archivo.** Permite unir los casos o las variables de dos archivos distintos. El sistema le pedirá un archivo de disco para fundir con el que tiene abierto. A través de un cuadro de diálogo, se indicarán las variables de los archivos con las que usted se quedará. Las variables coincidentes se asignarán automáticamente al nuevo archivo (aparecerán en el panel de la derecha) y las que no, vendrán indicadas con (+) o con (*) en el panel de la izquierda, según el archivo al que pertenezcan. Desde este panel se podrán incluir variables en el nuevo archivo, bien de forma independiente, bien agrupadas por parejas.
- **Agregar datos.** Combina grupos de casos (de resumen únicos) y crea un nuevo archivo de datos agregados (llamado por defecto, AGR.SAV). Se deben indicar las variables por las que se va a segmentar el archivo y las variables que aparecerán agrupadas. El nuevo archivo de datos contiene un caso para cada grupo.

- **Diseño ortogonal.** La generación de diseño ortogonal crea un archivo de datos que contiene un diseño de efectos principales ortogonales que permite contrastar, de forma estadística, varios factores sin probar cada combinación de los niveles de factores. Este diseño puede mostrarse mediante Mostrar diseño. Se debe definir al menos un factor, cuyo nombre puede ser cualquier nombre de variable, y definir los valores de cada uno de ellos. Si no se reemplaza el actual archivo de trabajo, se crea automáticamente un archivo en disco llamado por defecto ORTO.SAV, que contiene el diseño ortogonal.
- **Segmentar archivo.** Divide el archivo de datos en distintos grupos para el análisis, basándose en los valores de una o más variables de agrupación. Si selecciona varias variables de agrupación, los casos se agruparán por variable dentro de las categorías de la variable anterior de la lista Grupos basados en. Cuando pase al análisis obtendrá resultados divididos, según los valores de las variables que haya seleccionado con la opción Organizar los resultados por grupos.

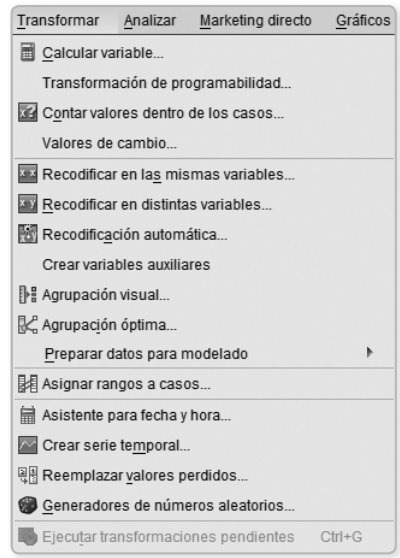



Si elige comparar los grupos, los resultados se mostrarán juntos según los grupos, para poder compararlos. Es necesario que el archivo esté ordenado según las variables de agrupación; de lo contrario, tendrá que indicárselo al sistema.

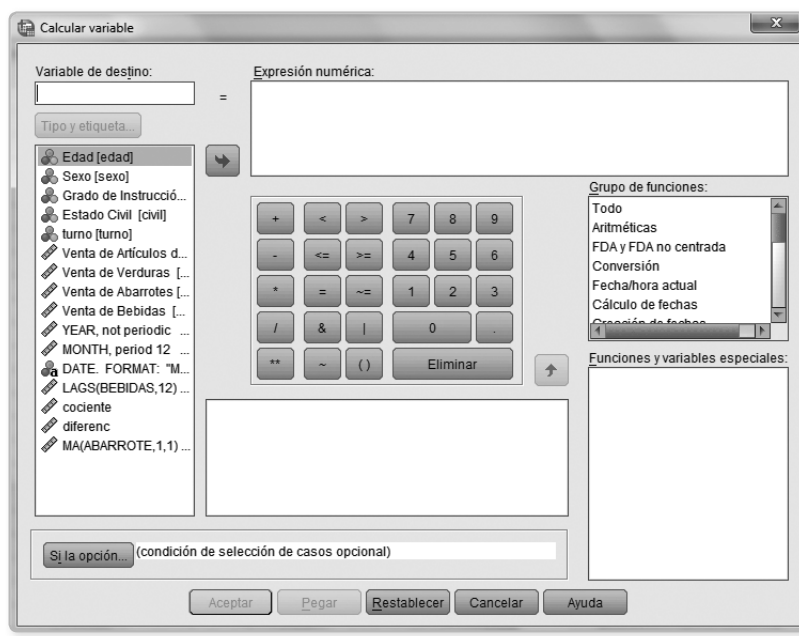
- **Seleccionar casos.** Proporciona varios métodos para seleccionar un subgrupo de casos, basándose en criterios que incluyen variables y expresiones complejas. Asimismo, se puede seleccionar una muestra aleatoria de casos. Los criterios usados para definir un subgrupo pueden incluir valores y rangos de las variables (del tipo que sean), números de caso (filas), expresiones aritméticas y lógicas, además de funciones. Los casos no seleccionados debido a que no cumplen los criterios pueden ser filtrados o eliminados. Con este procedimiento se consigue que cualquier análisis posterior se realice únicamente con los casos seleccionados.
- **Ponderar casos.** Permite asignar un peso o ponderación a cada caso según la variable de frecuencia que indiquemos. Es decir, esta variable debe indicar el número de observaciones únicas representadas por un único caso. Los casos con valores perdidos, negativos o cero para la variable de ponderación se excluyen del análisis, pero sí se computan los valores fraccionarios (en las tablas de contingencia se redondearía al entero más próximo). La variable de ponderación aplicada se mantiene activa hasta que se seleccione otra o se desactive la ponderación. Si se guarda el archivo, la información de ponderación se guardará con él.

### E. Menú Transformar (Transform)

Las funciones importantes del menú Transformar son las siguientes:




- **Calcular (*Compute*).** La opción Calcular permite asignar a una variable nueva o ya existente una expresión numérica o alfanumérica, la misma que puede contener operaciones y funciones matemáticas preincorporadas con números y variables existentes. Es una forma inmediata de crear nuevas variables con valores constantes, pero también le permite seleccionar los casos a los que aplicar la operación, según las condiciones que establezca.



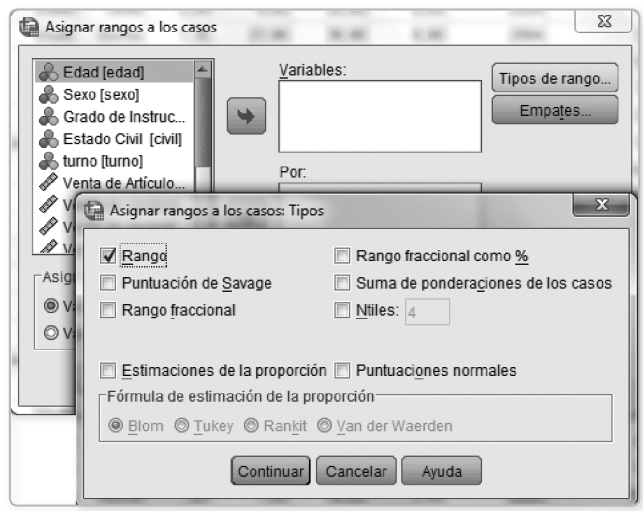
- **Aleatorización.** La semilla de aleatorización se establece cada vez que se inicia una sesión. Esta semilla es utilizada por el generador de números pseudoaleatorios para los procedimientos que lo precisen, y se regenera cada vez que se necesita. La opción Semilla de aleatorización le permite establecer la semilla de forma que se pueda repetir una secuencia previamente generada de números.

- **Contar números.** La función Contar apariciones, como su propio nombre indica, cuenta las veces que se repite un determinado valor o conjunto de valores en más de una variable. Dicho de otra forma, cuenta las veces que se repite un valor en un mismo caso. El cuadro de diálogo solicita la variable de destino (nueva o existente) donde se almacenará el número de veces que aparece el valor. Debe indicar las variables donde contar las apariciones, definir el valor o rango de valores que se pretende contar y hasta puede realizar una selección previa de casos, según determinadas condiciones (*Si...*).

- **Recodificar.** Esta opción, en las mismas o distintas variables, reasigna los valores de las variables existentes. Para una o más variables (del mismo tipo) debe definir el valor o rango de valores antiguos y el cambio directo (ya que no hay operación intermedia), por otros valores nuevos. El cuadro de diálogo permite especificar los casos a tratar.

- **Categorizar variable.** La función Categorizar variables convierte datos numéricos continuos en un número discreto de categorías, que se almacenan en nuevas variables. Los datos se categorizan según grupos percentiles, por lo que cada grupo contiene aproximadamente el mismo número de valores diferentes. Por ejemplo, la especificación de cuatro grupos asignará el valor 1 a los casos situados bajo el percentil 25; valor 2 a los casos entre el percentil 25 y el 50; valor 3 a los casos situados entre el percentil 50 y el 75, y valor 4 a los casos por encima del percentil 75.

- **Asignar rangos a casos.** La opción Asignar rangos a casos crea nuevas variables que contienen rangos, puntuaciones de Savage y normales, y los valores de los percentiles para las variables numéricas. Por defecto, se crean tantos rangos como casos existen: el menor elemento sería el 1 y el mayor, el número de casos; los empates se resuelven con la media de los valores del rango.



Los nombres de las nuevas variables, así como sus etiquetas descriptivas, se generan automáticamente en función del nombre de la variable original y de las medidas seleccionadas. Además, se crea una tabla resumen con la lista de las variables originales, nuevas variables y etiquetas de variable. Los rangos se pueden establecer en orden ascendente o descendente. Se puede indicar más de una variable, con lo que el rango se calcularía para cada combinación de las mismas. La lista Por define variables de agrupación, con lo que el procedimiento se ejecutaría independientemente para cada grupo definido por una o más variables indicadas. Se puede establecer uno o varios Tipos de rango que crearán sus respectivas variables. Por último, los empates pueden dirimirse con la media (por defecto), mayor, menor o con valores secuenciales.

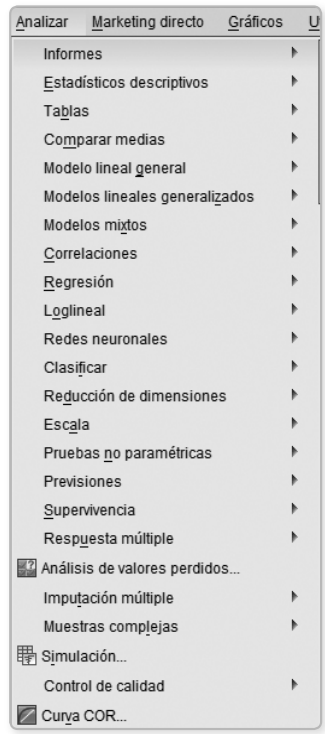
- **Recodificación automática.** Convierte los valores numéricos y de cadena en valores.
- **Enteros consecutivos.** La recodificación se realiza sobre una nueva variable que conserva las etiquetas de valor de la variable antigua. Para los valores que no tienen una etiqueta de valor ya definida, se utiliza el valor original como etiqueta del valor recodificado. Además, se genera una tabla en el Visor, que muestra los valores antiguos, los nuevos y las etiquetas de valor. La recodificación puede realizarse en orden ascendente o descendente.; los valores de cadena se recodifican por orden alfabético, con las mayúsculas antes que las minúsculas y los valores perdidos se recodifican como valores perdidos mayores que cualquier valor no perdido, conservando el orden. Por ejemplo, si la variable original posee 10 valores no perdidos, el valor perdido mínimo se recodificará como 11, y el valor 11 será un valor perdido para la nueva variable.
- **Crear serie temporal.** La opción Crear serie temporal crea nuevas variables basadas en funciones de variables de series temporales numéricas existentes. Una **serie temporal** es una secuencia de valores de una variable, donde cada caso representa una observación en un momento de tiempo distinto. Los valores obtenidos al crear una serie de este tipo son útiles en muchos procedimientos de análisis de series temporales. Existen varias funciones disponibles, como la diferencia, medias móviles, medianas móviles, retardo y adelanto. Por ejemplo, la diferencia asignaría a cada caso de la nueva variable la diferencia entre el valor del mismo caso y del caso anterior de la variable original.

  El sistema genera una tabla en el Visor con información sobre el procedimiento realizado, y los nombres de las nuevas variables se crean automáticamente sobre la base de la variable usada para crearlas, y conservan cualquier etiqueta de valor ya definida. Algunas medidas de series temporales no se pueden calcular si hay valores perdidos en la serie, y pueden causar problemas en determinados análisis que necesitan la existencia de todas las observaciones. Para solventar este problema se tiene el procedimiento.
- **Reemplazar valores perdidos.** Esta opción permite Reemplazar valores perdidos, que crea nuevas variables de series temporales a partir de las existentes, pero reemplazando los valores perdidos por estimaciones calculadas. El método de estimación puede ser la media de los valores existentes (por defecto), que asignaría la media obtenida a todos los valores perdidos, la media o mediana de puntos adyacentes, la interpolación lineal y la tendencia lineal en un punto. De la misma forma que en el procedimiento anterior, el sistema genera una tabla de información y los nombres de las nuevas variables se crean automáticamente. conservándose cualquier etiqueta de valor definida.

  Las transformaciones que haga se ejecutarán automáticamente, salvo que se haya indicado lo contrario, mediante las opciones generales del programa. Una vez en este panel se selecciona la pestaña Datos y se marca la opción Calcular los valores antes de usarlos. Las transformaciones generadas con los procedimientos calcular o recodificar, entre otros (en general, los que no requieren una lectura diferente de los datos), se van acumulando sin ser ejecutadas hasta que se lo indique mediante el comando.
- **Ejecutar transformaciones pendientes.** Esto puede ser útil trabajando con archivos grandes para retrasar la ejecución y ganar tiempo de procesamiento.

### F. Menú Analizar (Analyze)

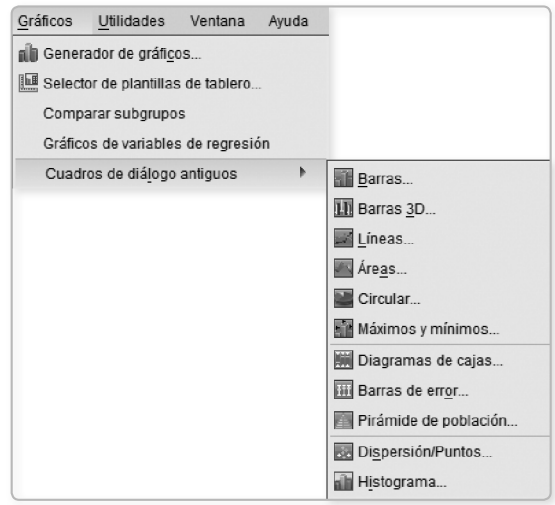
Este menú permite obtener resultados de las diferentes técnicas estadísticas que se desea utilizar para solucionar el problema planteado.




Es el menú más importante, porque desde aquí se indica el tipo de análisis que se quiere realizar con los datos. Cada fila (terminada en una punta de flecha) ofrece a su vez otros menús, cada uno con su correspondiente cuadro de diálogo, que es el lugar donde se especifica el número de variables que entran en el análisis, además de otra serie de detalles que se verán más adelante. Entre las técnicas estadísticas más usadas están los Estadísticos descriptivos, Correlaciones, Regresión, etc.

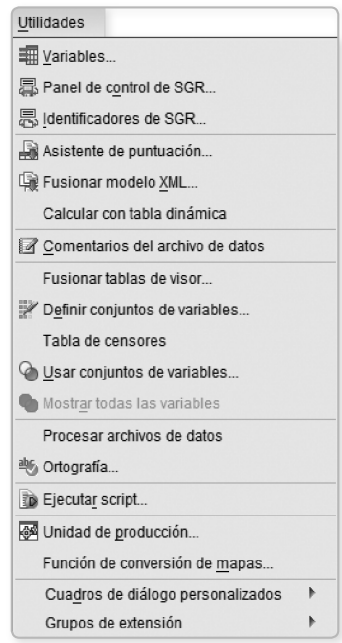
### G. Menú Gráficos (Graphs)

El menú Gráficos permite realizar una gran cantidad de gráficos orientados al análisis de los datos, como barras en 3D (simples, en grupos, apiladas), barras (simples, en grupos, apiladas), líneas (simples, múltiple, líneas de caída, proyección), aéreas (simples, apiladas), sectores (simples separados, máximo–mínimo, máximo–mínimo cierre), línea de diferencia, diagrama de cajas (simples, en grupos), barras de intervalo (simples, en grupos), barras de error (simples, en grupos, de los gráficos de barras, líneas, áreas, intervalos de confianza, SD, SE), etc.




### H. Menú Utilidades (Utilities)

Cuando se necesita información detallada sobre las variables se dispone de utilidades.




- **Variables.** Muestra un cuadro dividido en dos paneles: uno contiene la lista de variables y otro presenta información sobre la que esté seleccionada.
- **Información de variables.** Información del archivo que muestra, a través del Visor de resultados, un informe más o menos completo sobre las variables que contiene el archivo de datos que se está usando.

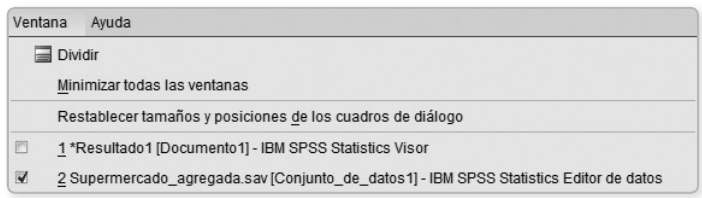
- **Definición conjunta de variables.** Permite definir los nombres de los conjuntos y las variables que se asigna a cada uno. Los nombres tienen un máximo de doce caracteres, incluidos los espacios, y no distingue entre mayúsculas y minúsculas. Posteriormente, cuando se quiera utilizarlos, se irá a la opción.
- **Usar conjuntos.** Use conjuntos y asígnelos. Por defecto, se utilizan dos conjuntos predefinidos por el sistema: ALLVARIABLES (contiene todas las variables del archivo de datos, incluidas las nuevas creadas durante la sesión) y NEWVARIABLES (solo contiene las creadas durante la sesión). Estos conjuntos pueden quitarse de la lista y situar otros.

### I. Complementos (Add ons)

Se integra con todas las funciones, como el proceso de análisis, planificación, recopilación de datos, acceso a datos, gestión de datos y la preparación, análisis de datos, presentación de informes y el despliegue etc.

### J. Ventana (Windows)

Permite disponer, seleccionar y controlar los atributos de las diferentes ventanas. También permite gestionar las ventanas abiertas y cambiar de una ventana a otra.




### K. Menú Ayuda (Help)

Por último, el menú que aparece con la expresión **?** se refiere a la ayuda que ofrece el programa, y será muy útil para el usuario en todo momento. Puede accederse a él desde la ventana Editor de datos, Visor de resultados y, en ambos casos, ofrece el siguiente aspecto:




Permite obtener todos los temas expuestos en el formato de ayuda de Windows. Uno de ellos se expone a modo de breve tutorial. Otras opciones de la ayuda son la ***Guía de sintaxis***, para aprender la sintaxis de los comandos internos; y el ***Asesor estadístico,*** que le guiará a través de una serie de pasos hasta obtener los resultados que desee.

Otra forma de obtener ayuda es de manera contextual. Por ejemplo, en los cuadros de diálogo, pulsando el botón derecho del mouse, se muestra una breve descripción del elemento sobre el que se halla, o bien se muestra un menú contextual con la opción ¿Qué es esto? Esta posibilidad está incluso en los elementos gráficos en la ventana Visor. Desde Temas también puede escribirse un tipo de análisis (por ejemplo: Frecuencias) y se muestra una descripción de sus opciones. Asimismo, es posible obtener ayuda pulsando el botón derecho del mouse. Por ejemplo, si se encuentra en el cuadro de diálogo donde se calculan los estadísticos descriptivos y se quiere saber la definición de curtosis, solo necesita seleccionarlo y pulsar el botón derecho del mouse. Por último, en muchos cuadros un botón de ayuda aparece para acceder directamente al tema referente.

### 2.2.2. USO DE LOS CUADROS DE DIÁLOGO

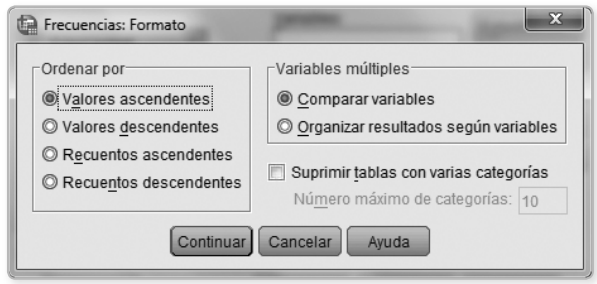
Antes de continuar, se debe conocer el uso de los cuadros de diálogo utilizados para seleccionar variables y opciones.

#### A. Selección de variables y opciones

Casi todos los procedimientos muestran este tipo de cuadros de diálogos como método visual y compacto de personalización de la tarea a ejecutar. Un cuadro de diálogo tiene el siguiente aspecto:




En la lista de la izquierda aparecen las variables susceptibles de ser empleadas en el procedimiento. Para seleccionarlas, pinche primero sobre ellas y después sobre el botón. Para seleccionar varias, si el procedimiento lo permite, use la tecla Ctrl. En la figura mostrada se han seleccionado las variables Frecuencia de..., Edad, de forma que ahora se puede pasar a la lista de la derecha. Además, pueden aparecer varias listas de destino, según las necesidades del procedimiento. A la izquierda de las variables se muestra un símbolo indicativo de su naturaleza (símbolo de color amarillo), que puede ser variable numérica de cualquier tipo, variable de cadena menor o igual a ocho caracteres, o variable de cadena mayor de ocho caracteres. Un cuadro puede contener diferentes controles (casillas de selección, listas desplegables, etc.) para configurar el análisis libremente. En la figura anterior, solo aparece uno de ellos (una casilla de selección), para indicar que se quiere ver las tablas de frecuencias. Cuando las opciones de configuración son demasiado extensas, se proporcionan otros cuadros de diálogo auxiliares accesibles desde determinados botones, mostrados normalmente en la parte inferior. Por ejemplo, pulsando el botón Formato, se presenta un nuevo cuadro:



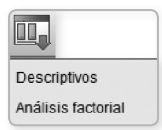

En este cuadro se podrá decidir la ordenación y otros parámetros del análisis. Cuando se haya terminado solo es preciso pulsar el botón Continuar para volver al cuadro de diálogo inicial. A veces, es preciso utilizar un cuadro de introducción de expresiones que puede contener variables, operaciones, constantes, funciones, etc. Considere que las constantes de cadena deben ir entre comillas o apóstrofes, y que las constantes numéricas se deben escribir en formato americano, con el punto como separador de la parte decimal.

Por último, la parte extrema derecha del cuadro (a veces, la parte inferior) muestra un conjunto de botones comunes a casi todos los cuadros:

- **Aceptar:** Ejecuta el procedimiento.
- **Pegar:** Pega en la ventana de sintaxis la operación que se está realizando.
- **Restablecer:** Restablece todas las opciones con sus valores por defecto.
- **Cancelar:** Cancela el procedimiento.
- **Ayuda:** Acceso directo a la ayuda sobre ese procedimiento.

### B. Rellamada de los cuadros de diálogo

En la barra de herramientas se encuentra un botón para acceder rápidamente a los cuadros de diálogo usados con mayor frecuencia. Al pulsar este botón, se muestra una lista ordenada de manera que los procedimientos usados más recientemente se situarán en posiciones superiores:



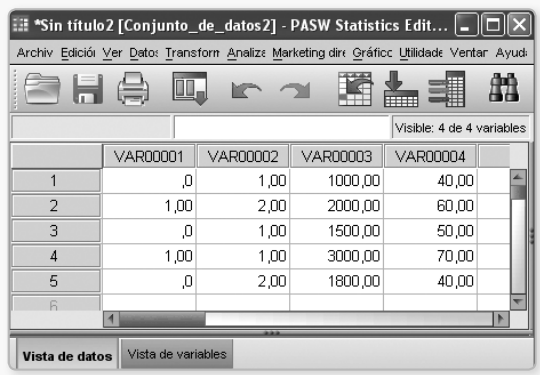

Al terminar la sesión, esta lista se almacenará en disco para poder usarla en sesiones posteriores. De esta manera, los procedimientos que se usen con elevada frecuencia siempre estarán rápidamente disponibles.

## Ejercicio práctico 1: Manipulación de datos

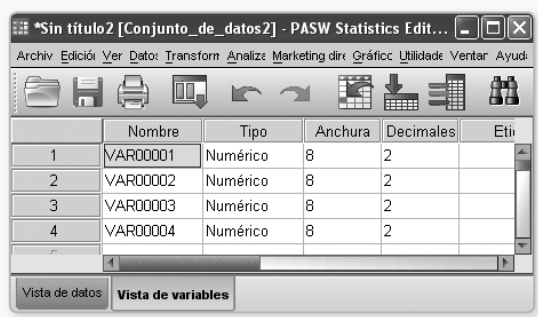
Los datos corresponden a 5 sujetos (casos) medidos en cuatro variables: Sexo (0: mujer y 1: varón), Nivel cultural (1: bajo y 2: alto), Ingresos en nuevos soles y Peso en kilogramos.

| Sujetos | Sexo | Nivel cultural | Ingresos | Peso |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 000 | 40 |
| 2 | 1 | 2 | 2 000 | 60 |
| 3 | 0 | 1 | 1 500 | 50 |
| 4 | 1 | 1 | 3 000 | 70 |
| 5 | 0 | 2 | 1 800 | 40 |

Si bien es cierto lo más recomendable es definir las variables, debe estar atento a lo que sucede si primero se ingresan los datos a la matriz. Se elabora una matriz de cinco filas por cuatro columnas; los datos se introducen colocando un caso (sujeto) en cada fila y una variable en cada columna. Para ello, simplemente se escriben las puntuaciones correspondientes a cada caso. Para moverse de una celda a otra, se pulsan las teclas marcadas con flechas. En el siguiente ejemplo, los datos de los cinco primeros sujetos se colocan del siguiente modo:




En la figura anterior se observa que el SPSS define los nombres de las variables por defecto, como: VAR00001, VAR00002, VAR00003 y VAR00004. Asimismo, se han definido los valores numéricos con dos decimales.

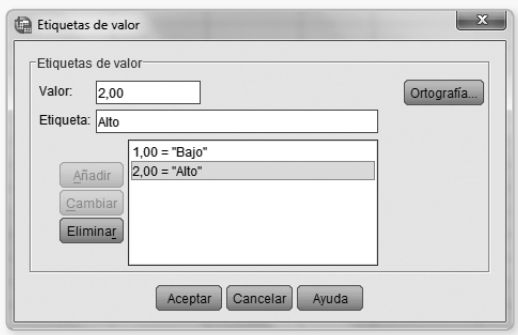


Una vez introducidos los datos en la matriz de datos, es necesario dar nombre a cada una de las variables. Si se desea cambiar el nombre de una variable y definir sus características (nombre, tipo, ancho de columna, decimales, etiquetas de valores, formato de columna, definición de valores perdidos, etc.) pulse con el mouse en la solapa Vista de variables (variable View). A continuación, se muestra una ventana que contiene tantas filas como variables hay definidas en el Editor de datos. En este caso, son cuatro filas porque existen cuatro variables (Sexo, Nivel cultural, Ingresos y Peso).

En cada fila se da la posibilidad de modificar los parámetros, como ***nombre*** (*Name*) de la variable (teniendo en cuenta que no puede tener más de ocho caracteres de largo); además de otras propiedades como ***tipo*** (*Type*: numérico, cadena, etc.), ***anchura*** (*Width*), el número de ***decimales*** (*Decimals*), la ***etiqueta*** (*Label*), el significado de los ***valores*** (*Values*) que admite, los valores ***perdidos***, la ***alineación*** del texto, etc. Solo se necesita posicionar el cursor en cada casilla (celda) y escribir la información deseada.

| | Nombre | Tipo | Anchura | Decimales | Etiqueta | Valores | Perdidos | Columnas | Alineación | Medida | Rol |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | sexo | Numérico | 8 | 2 | Sexo | Ninguna | Ninguna | 8 | Derecha | Escala | Entrada |
| 2 | nivel | Numérico | 8 | 2 | Nivel Cultural | Ninguna | Ninguna | 8 | Derecha | Escala | Entrada |
| 3 | ingresos | Numérico | 8 | 2 | Ingresos | Ninguna | Ninguna | 8 | Derecha | Escala | Entrada |
| 4 | peso | Numérico | 8 | 2 | Peso | Ninguna | Ninguna | 8 | Derecha | Escala | Entrada |
| 5 | | | | | | | | | | | |

En este ejemplo, los nombres (Names) de las variables son: *sexo, nivel, ingresos* y *peso*, respectivamente. Algunas de las casillas de la ventana Vista de variables incorporan su propio cuadro de diálogo que se activa al seleccionarlas. Por ejemplo, al pulsar en Valores, en la variable Nivel cultural, aparece lo siguiente:




Desde este cuadro de diálogo, se puede dar etiqueta a los valores de una variable. Por ejemplo, en Nivel cultural, el valor 1 significa nivel cultural bajo y el valor 2, nivel cultural alto. Cada vez que se introduzca la etiqueta de un valor se pulsa la tecla Añadir (*Add*). Para terminar, pulse Aceptar (*OK*).

Una vez definidos todos los parámetros de las variables, se puede volver a visualizar los datos numéricos; para esto, se pincha en la solapa Vista de datos. En la matriz de datos que se muestra a continuación se puede observar que las variables ya están definidas.

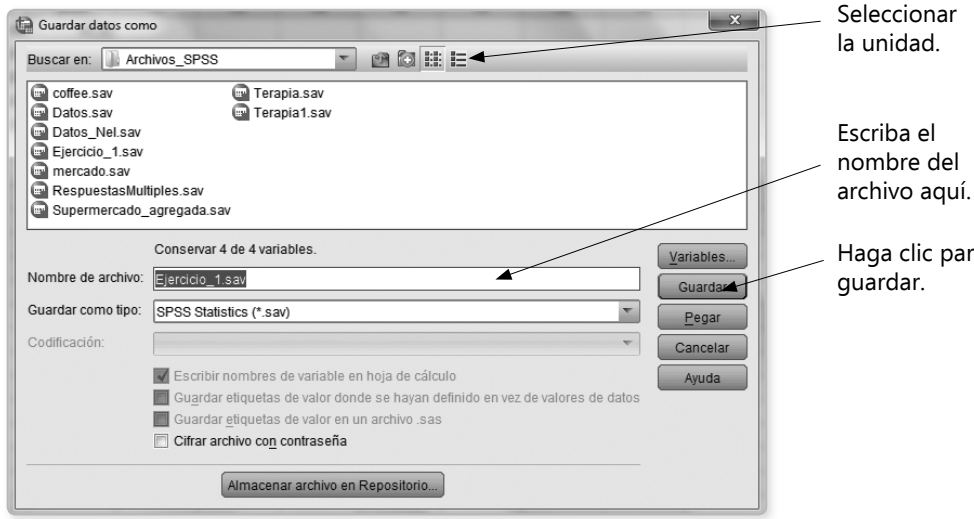
| | sexo | nivel | ingresos | peso | var |
|---|---|---|---|---|---|
| 1 | 0,00 | Bajo | 1000,00 | 40,00 | |
| 2 | 1,00 | Alto | 2000,00 | 60,00 | |
| 3 | 0,00 | Bajo | 1500,00 | 50,00 | |
| 4 | 1,00 | Bajo | 3000,00 | 70,00 | |
| 5 | 0,00 | Alto | 1800,00 | 40,00 | |

Data View | Variable View

Algo importante que se debe realizar antes de continuar, es guardar la información en un archivo de disco.

### C. Guardar

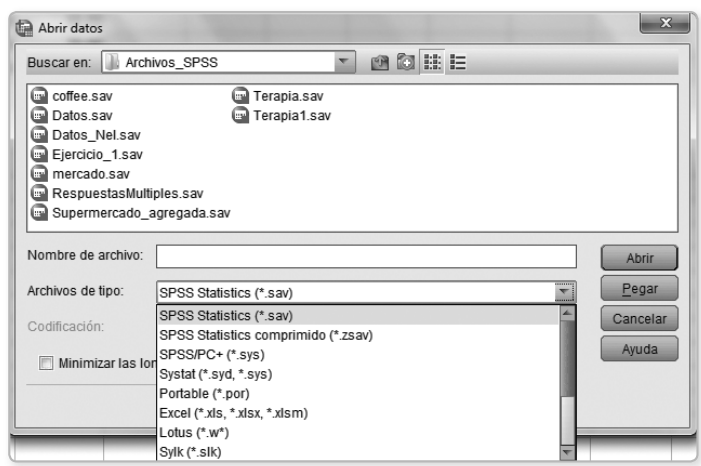
Para guardar la información introducida en el Editor de datos, se selecciona el menú Archivo de la barra de herramientas, se elige Guardar y aparece el correspondiente cuadro de diálogo.




El SPSS guarda archivos de datos con la extensión ***sav***; por tanto, solo se necesita dar un nombre a los datos (*ejercicio_1)* y, por defecto, se grabarán en un archivo con la extensión *sav* (aquí se le ha llamado *a*:\ *ejercicio_1.sav*). También es posible guardar archivos de resultados; para ello, sitúese en la ventana Visor de resultados, seleccione el menú Guardar y aparecerá un cuadro de diálogo idéntico al de la figura anterior. El procedimiento a seguir es exacto a archivar datos, con la única diferencia que la extensión de este tipo de archivos es ***spo***; por tanto, los archivos SPSS con la extensión *sav* contienen datos y los de la extensión *spo,* resultados.

### D. Abrir archivos de datos existentes

Cuando el archivo de datos se encuentra ya introducido en el ordenador o en un disquete, solo se necesitará abrirlo. Para ello, seleccione el menú Abrir y la unidad donde se encuentra el archivo, y entonces se mostrará el cuadro de diálogo correspondiente a Abrir. Sus datos se encuentran grabados en el archivo *a*:*practica.sav*; para acceder a ellos, selecciónelo y haga clic en el botón Abrir.



Una vez completada esta operación, aparecerán los datos en el Editor de datos. A continuación, se muestra la matriz para los cinco sujetos en vista Datos, con sus respectivas variables (sexo, nivel, ingresos y peso).




| | sexo | nivel | ingresos | peso | var |
|---|---|---|---|---|---|
| 1 | Mujer | Bajo | 1000,00 | 40,00 | |
| 2 | Varòn | Alto | 2000,00 | 60,00 | |
| 3 | Mujer | Bajo | 1500,00 | 50,00 | |
| 4 | Varòn | Bajo | 3000,00 | 70,00 | |
| 5 | Mujer | Alto | 1800,00 | 40,00 | |
| 6 | | | | | |

**NOTA**

*Siempre que se desee iniciar una sesión con el SPSS donde se lleve a cabo uno o más análisis de datos, es necesario que el archivo de datos se encuentre abierto en el Editor de datos del SPSS. Para que el SPSS realice algún tipo de análisis estadístico, no se requiere que los datos estén grabados en un archivo, sino que dicho archivo esté abierto.*

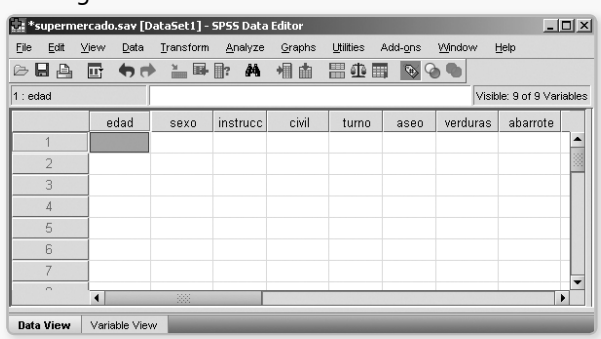
## Ejercicio práctico 2: Manipulación de datos

Si desea llenar la base de datos de la encuesta supermercado, primero defina los parámetros para las variables *edad, sexo, grado de instrucción, estado civil, horario de compra (turno), venta de artículos de aseo, verduras abarrotes y bebida*. En cada fila se escribirá el ***nombre*** (name) de las variables, el ***tipo*** (numérico), la ***anchura*** (8 caracteres), el número de ***decimales*** (2 decimales), la ***etiqueta*** (*Label*), el significado de los ***valores*** que admite (tener presente que las variables edad, sexo, grado de instrucción, estado civil y turno tienen valores), los valores ***perdidos***, la ***alineación*** del texto, etc.

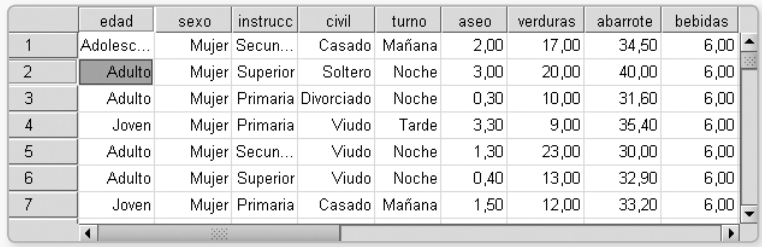
Para escribir la información de cada variable simplemente coloque el cursor en cada casilla y escriba la información deseada.

| | Nombre | Tipo | Anchura | Decimales | Etiqueta | Valores | Perdidos | Columnas | Alineación | Medida | Rol |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | edad | Numérico | 8 | 0 | Edad | {0, Adolesc… | Ninguna | 6 | Derecha | Nominal | Entrada |
| 2 | sexo | Numérico | 8 | 0 | Sexo | {0, Mujer}… | Ninguna | 6 | Derecha | Nominal | Entrada |
| 3 | instrucc | Numérico | 8 | 0 | Grado de Instru… | {1, No Tiene… | Ninguna | 5 | Derecha | Nominal | Entrada |
| 4 | civil | Numérico | 8 | 0 | Estado Civil | {1, Soltero}… | Ninguna | 6 | Derecha | Nominal | Entrada |
| 5 | turno | Numérico | 8 | 0 | turno | {0, Mañana}… | Ninguna | 5 | Derecha | Nominal | Entrada |
| 6 | aseo | Numérico | 8 | 2 | Venta de Artícu… | Ninguna | Ninguna | 5 | Derecha | Escala | Entrada |
| 7 | verduras | Numérico | 8 | 2 | Venta de Verdu… | Ninguna | Ninguna | 6 | Derecha | Escala | Entrada |
| 8 | abarrote | Numérico | 8 | 2 | Venta de Abarr… | Ninguna | Ninguna | 6 | Derecha | Escala | Entrada |
| 9 | bebidas | Numérico | 8 | 2 | Venta de Bebidas | Ninguna | Ninguna | 6 | Derecha | Escala | Entrada |

Una vez definidas todas las variables de la encuesta supermercados, diríjase a la pestaña Editor de datos, cuyo aspecto es el siguiente:



Los datos se introducen colocando un caso (sujeto) en cada fila y una variable en cada columna. Para ello, simplemente se escriben las puntuaciones correspondientes a cada caso. Para moverse de una celda a otra, se pulsan las teclas marcadas con flechas. En este ejercicio, los datos están representados por 30 sujetos y cada sujeto tiene nueve variables; es decir, la matriz de datos será de 30 filas por 9 columnas, tal como se aprecia en el siguiente gráfico:



| | edad | sexo | instrucc | civil | turno | aseo | verduras | abarrote | bebidas |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Adolesc... | Mujer | Secun... | Casado | Mañana | 2,00 | 17,00 | 34,50 | 6,00 |
| 2 | Adulto | Mujer | Superior | Soltero | Noche | 3,00 | 20,00 | 40,00 | 6,00 |
| 3 | Adulto | Mujer | Primaria | Divorciado | Noche | 0,30 | 10,00 | 31,60 | 6,00 |
| 4 | Joven | Mujer | Primaria | Viudo | Tarde | 3,30 | 9,00 | 35,40 | 6,00 |
| 5 | Adulto | Mujer | Secun... | Viudo | Noche | 1,30 | 23,00 | 30,00 | 6,00 |
| 6 | Adulto | Mujer | Superior | Viudo | Noche | 0,40 | 13,00 | 32,90 | 6,00 |
| 7 | Joven | Mujer | Primaria | Casado | Mañana | 1,50 | 12,00 | 33,20 | 6,00 |

## Ejercicios propuestos: Crear archivos

**Ejercicio 1.** La TABLA 1 contiene indicadores económicos, demográficos, sanitarios y de otros tipos para diversos países del mundo.

**Tabla 1**

| país | poblac | relig | espvidaf | espvidam | región | fertilid | clima |
|---|---|---|---|---|---|---|---|
| Alemania | 81200 | Protestante | 79 | 73 | 1 | 1,5 | 7 |
| Armenia | 3700 | Ortodoxa | 75 | 68 | 5 | 3,2 | N.D. |
| Brasil | 156600 | Católica | 67 | 57 | 6 | 2,7 | 4 |
| China | 1205200 | Taoista | 69 | 67 | 3 | 1,8 | 7 |
| Dinamarca | 5200 | Protestante | 79 | 73 | 1 | 1,7 | 7 |
| Egipto | 60000 | Musulmana | 63 | 60 | 5 | 3,8 | 1 |
| España | 39200 | Católica | 81 | 74 | 1 | 1,4 | 7 |
| Gambia | 959 | Musulmana | 52 | 48 | 4 | 6,3 | 4 |
| Grecia | 10400 | Ortodoxa | 80 | 75 | 1 | 1,5 | 7 |
| Hong Kong | 5800 | Budista | 80 | 75 | 3 | 1,4 | 4 |
| India | 911600 | Indú | 59 | 58 | 3 | 4,5 | 5 |
| Letonia | 2700 | Protestante | 75 | 64 | 2 | 2,0 | 6 |
| Libia | 5500 | Musulmana | 65 | 62 | 5 | 6,4 | 5 |
| México | 91800 | Católica | 77 | 69 | 6 | 3,2 | 3 |
| Portugal | 10500 | Católica | 78 | 71 | 1 | 1,5 | 6 |
| Rusia | 149200 | Ortodoxa | 74 | 64 | 2 | 1,8 | 8 |
| Senegal | 8700 | Musulmana | 58 | 55 | 4 | 6,1 | 4 |
| Sudáfrica | 43900 | N.D. | 68 | 62 | 4 | 4,4 | 2 |
| Suecia | 8800 | Protestante | 81 | 75 | 1 | 2,1 | 8 |
| Taiwán | 20944 | Budista | 78 | 72 | 3 | N.D. | N.D. |

*Fuente: Técnicas estadísticas con SPSS, Prentice Hall.*

a. Abra un archivo de datos llamado «Países», de forma que contenga la información recogida en la tabla 1, y defina las variables tal como sigue:

| Variable | Tipo | Etiqueta | Anchura / N° Caracteres | Alineación |
|---|---|---|---|---|
| país | Cadera | País | 16 | Izquierda |
| poblac | Punto | Población x 1000 | 8 | Derecha |
| relig | Cadena | Religión mayoritaria | 8 | Izquierda |
| espvidaf | Numérico | Esperanza de vida femenina | 4 | Centrado |
| espvidam | Numérico | Esperanza de vida masculina | 5 | Centrado |
| región | Numérico | Región económica | 12 | Derecha |
| fertilid | Numérico | Número promedio de hijos | 6 | Derecha |
| clima | Numérico | Clima predominante | 3 | Derecha |

b. Los valores de la variable **fertilid** tienen un decimal; mientras que el resto son sin decimales.

c. Establezca etiquetas de valor para la variable **región**:

1: OCDE
2: Europa Oriental
3: Asia / Pacífico
4: África
5: Oriente Medio
6: América Latina

d. Establezca etiquetas de valor para la variable **clima**:

1: Desierto
2: Árido / Desierto
3: Árido
4: Tropical
5: Mediterráneo
6: Marítimo
7: Templado
8: Ártico/Templado

e. Considere los valores ND (datos no disponibles) como valores perdidos del sistema.

f. Considere las regiones de Asia/Pacífico y de América Latina perdidas por el usuario.

g. Guarde en el disquete el archivo de datos que ha creado.

**Ejercicio 2.** Una agencia de alquiler de coches está estudiando renovar parte de su parque automovilístico de todo terrenos. En la TABLA 1 se muestran las características más importantes de los modelos que se está planteando comprar.

**Tabla 1:** Principales características de los modelos de coches todo terreno, con posibilidades de ser comprados por la agencia de alquiler

| Marca | Modelo | PVP (en ptas.) | N° plazas | Consumo urbano | Velocidad máx. | Aceleración |
|---|---|---|---|---|---|---|
| JEEP | Wrangler 4.0 | 3.250.856 | 4 | 19,4 | 170 | 9,4 |
| FORD | Maverick 2.4 GL x 5p | 4.461.000 | 7 | 13,3 | 160 | N.D. |
| LAND ROVER | Defender 90 TDI t. L | 3.062.000 | 6 | 13,6 | N.D. | 16,8 |
| MERCEDES | GE 320 Corto | 9.349.120 | 5 | 17,7 | 175 | 12,0 |
| NISSAN | Terrano II 2.4 LX 3p | 3.134.000 | 5 | 13,1 | 160 | 13,2 |
| MITSUBICHI | Montero Largo 3.0 GL | 4.747.813 | 7 | 16,2 | 175 | 12,1 |

Nota: N.D. significa dato no disponible.

*Fuente: VISAUTA VINACUA, B. (1997). Análisis estadístico con SPSS para Windows. Volumen 1, Estadística Básica, McGraw-Hill.*

a. Abra un nuevo archivo de datos denominado «Terreno» que contenga la información recogida en la TABLA 1, considerando las especificaciones de las variables recogidas en la TABLA 2.
b. Defina la variable **marca** como una variable de tipo numérico sin decimales, que asigne una etiqueta de valor que se corresponda con un número del 1 al 6 para cada una de las distintas marcas de todoterreno.
c. Defina la variable **modelo** como una variable de tipo cadena de 20 caracteres.
d. Defina las variables **plazas, pvp** y **velocidad** como variables numéricas sin algún decimal.
e. Defina las variables **consurb** y **acelerac** como variables numéricas de ancho cinco y con un decimal.
f. Considere los valores no numéricos de las variables **velocidad** y **acelerac** como valores perdidos del sistema.
g. Considere como valores perdidos del usuario la velocidad máxima 170 km/h y la aceleración de 0 a 100.

**Tabla 2:** Nombre y etiquetas de las variables del archivo

| Texto Tabla 1 | Variable | Etiqueta |
|---|---|---|
| Marca | marca | Marca |
| Modelo | modelo | Modelo |
| PVP (en ptas.) | pvp | Precio (en ptas.) |
| N° de plazas | plazas | Número de plazas |
| Consumo urbano | consurb | Consumo urbano |
| Velocidad máx. | velocida | Velocidad máxima (en Km./h.) |
| Aceleración | acelerac | Aceleración de 0 a 100 |

CAPÍTULO

3

# MANIPULACIÓN DE CASOS Y VARIABLES

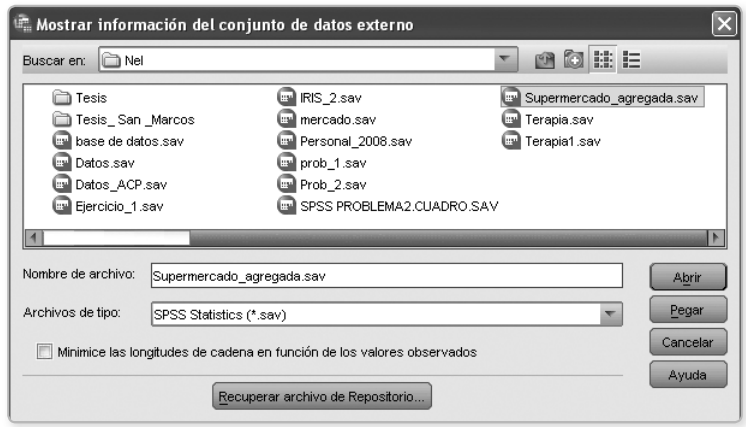
El objetivo principal de este capítulo es comprender cómo se manipulan los casos y variables, con la finalidad de hacer más fácil y práctico el análisis de datos. La manipulación puede consistir en obtener información relativa al archivo *.sav y las variables que lo componen, buscar un dato, insertar casos, insertar variables, ordenar casos, seleccionar casos, segmentar archivos, realizar cálculos con variables y recodificar variables.

## 3.1. VER INFORMACIÓN DE DATOS (DISPLAY DATA INFO…)

El primer paso en la manipulación de datos con SPSS es obtener información relativa al archivo *.sav y las variables que lo componen. Antes de abrir el archivo *.sav, ingrese al menú principal **Archivo**: **File** > **Diplay Data Info…** > **External File.**




De esta forma, se mostrará una ventana donde se tendrá que especificar el archivo.

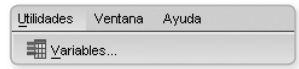



Luego, presione Aceptar. Toda la información se mostrará en la ventana de resultados (número de observaciones, nombres de variables, etiquetas, formato y rango de valores).

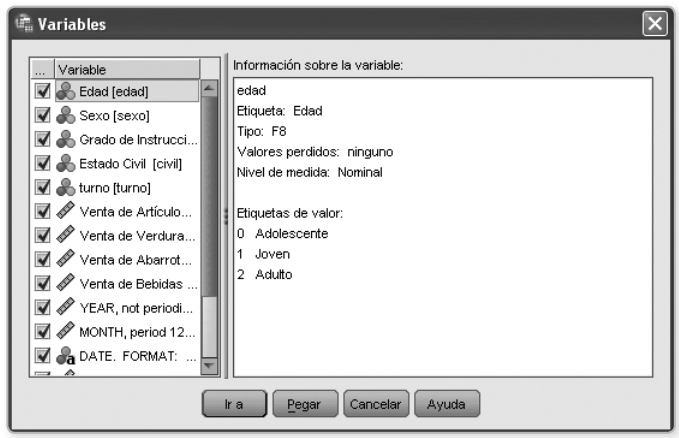
**Información de archivo**

| | | |
|---|---|---|
| Origen | | G:\Nel Quezada L\Archivos_SPSS\ Supermercado_agregada.sav |
| Tipo | | Archivo de datos de SPSS Statistics |
| Fecha de creación | | 17-MAY-2004 13:10:17 |
| Etiqueta | | Ninguna |
| Codificación de caracteres | | Desconocido |
| Contenido de archivo | Tipo de datos | Caso |
| | N.° de líneas de documentos | Ninguna |
| | Conjuntos de variables | Ninguna |
| | Predicción de información de fecha | Ninguna |
| | Definiciones de respuesta múltiple | Ninguna |
| | Información de Data Entry para Windows | Ninguna |
| | Información de TextSmart | Ninguna |
| | Información de modelador | Ninguna |
| Información de datos | N.° de casos | 30 |
| | N.° de elementos de variable definidos | 18 |
| | N.° de variables especificadas | 16 |
| | Variable de ponderación | Ninguna |
| | Comprimido | Sí |

## 3.2. VARIABLES ( )




Esta opción, accesible desde el ícono mostrado ubicado en el menú de herramientas estándar, permite obtener información de las variables del archivo (file) de trabajo activo. Luego, pulsando en el ícono (Botón) mostrado, resulta el siguiente cuadro de diálogo.




Este cuadro de diálogo permite observar información de todas las variables que intervienen en el archivo supermercado, como la variable Edad; si se quiere información de las otras variables, simplemente haga clic en la variable requerida y se mostrará la información similar a la variable Edad.

## 3.3. BUSCAR DATOS… (FIND)

Para buscar el valor específico de una variable, ubíquese en una celda cualquiera de la columna respectiva. Hay tres formas de alcanzar su objetivo: presionando el ícono Buscar datos, mediante el atajo Alt+F5 o a través del menú principal Edit > Find...




La ventana **Buscar datos en...** le pedirá el valor a buscar, y si la búsqueda se hace hacia adelante o hacia atrás, esta comenzará a partir de la celda seleccionada.

Es posible ignorar las minúsculas o mayúsculas presionando sobre el Control select respectivo. Se puede notar además, que la opción **Restringir búsqueda a casos filtrados** está desactivada, debido a que no se ha hecho una selección previa de casos.

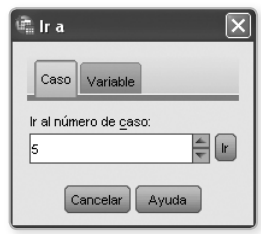



## 3.4. IR A CASO

Esta opción es accesible desde el ícono mostrado ubicado en el menú de herramientas estándar, o a través del menú principal **Data > Go to Case...**




Permite desplazarse al caso especificado por el número de fila de la tabla SPSS.




Luego de ingresar el caso especificado, se pulsa el botón Aceptar (OK) y se obtiene el resultado buscado.

## 3.5. INSERTAR VARIABLE (DATA > INSERT VARIABLE…)

Se puede insertar una variable a través del menú principal **Data > Insert variable...** o presionando el ícono mostrado en la barra de herramientas. SPSS insertará una variable (en blanco) en la posición en que se encuentre el cursor, desplazando todas las columnas hacia la derecha y nombrando dicha variable por defecto.

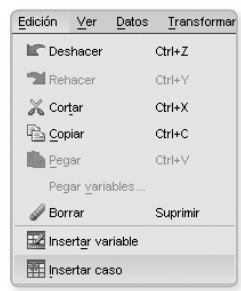
1 : VAR00001

| | edad | sexo | instrucc | civil | VAR00001 | turno | aseo | verduras | abarrote | bebidas |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Adolesc... | Mujer | Secun... | Casado | . | Mañana | 2,00 | 17,00 | 34,50 | 6,00 |
| 2 | Adulto | Mujer | Superior | Soltero | . | Noche | 3,00 | 20,00 | 40,00 | 6,00 |
| 3 | Adulto | Mujer | Primaria | Divorciado | . | Noche | ,30 | 10,00 | 31,60 | 6,00 |
| 4 | Joven | Mujer | Primaria | Viudo | . | Tarde | 3,30 | 9,00 | 35,40 | 6,00 |
| 5 | Adulto | Mujer | Secun... | Viudo | . | Noche | 1,30 | 23,00 | 30,00 | 6,00 |
| 6 | Adulto | Mujer | Superior | Viudo | . | Noche | ,40 | 13,00 | 32,90 | 6,00 |
| 7 | Joven | Mujer | Primaria | Casado | . | Mañana | 1,50 | 12,00 | 33,20 | 6,00 |
| 8 | Adolesc... | Mujer | Secun... | Casado | . | Mañana | 4,50 | 19,00 | 33,10 | 6,00 |
| 9 | Adulto | Varón | Superior | Divorciado | . | Noche | 2,50 | 18,00 | 35,60 | 6,00 |
| 10 | Joven | Varón | Primaria | Casado | . | Tarde | ,30 | 24,00 | 33,00 | 6,00 |
| 11 | Adulto | Varón | Superior | Soltero | . | Noche | 1,00 | 7,00 | 34,50 | 6,00 |
| 12 | Adulto | Mujer | Secun... | Casado | . | Noche | 6,00 | 10,00 | 33,20 | 6,00 |
| 13 | Adulto | Mujer | Superior | Viudo | . | Noche | 5,60 | 5,00 | 31,50 | 6,00 |
| 14 | Adulto | Varón | Superior | Viudo | . | Noche | 6,00 | 14,00 | 36,20 | 1,00 |
| 15 | Joven | Mujer | No Tie... | Soltero | . | Tarde | 1,20 | 15,00 | 36,80 | 1,00 |

En el ejemplo se puede observar que la variable insertada es VAR00001.

## 3.6. INSERTAR CASO (DATA > INSERT CASE…)

De manera análoga, dicha tarea se puede realizar a través del menú principal **Data** > **Insert Case...** o presionando el ícono mostrado en la barra de herramientas.

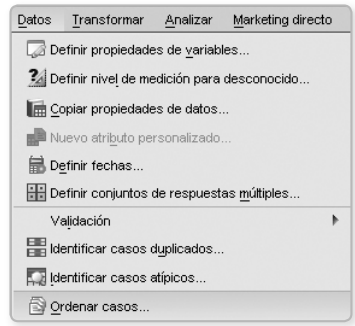



SPSS insertará un caso (en blanco) en la posición en que se encuentre el cursor, desplazando todas las filas hacia abajo.

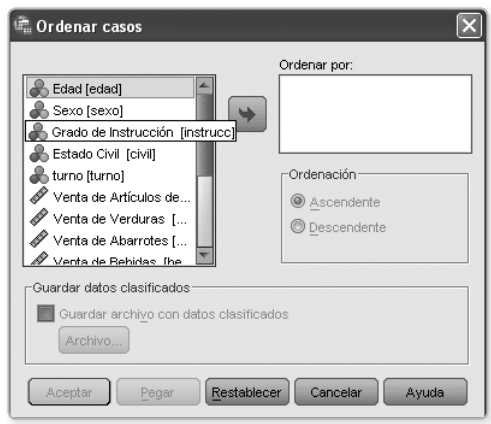
| | edad | sexo | instrucc | civil | turno | aseo | verduras | abarrote | bebidas |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Adolesc... | Mujer | Secun... | Casado | Mañana | 2,00 | 17,00 | 34,50 | 6,00 |
| 2 | Adulto | Mujer | Superior | Soltero | Noche | 3,00 | 20,00 | 40,00 | 6,00 |
| 3 | Adulto | Mujer | Primaria | Divorciado | Noche | ,30 | 10,00 | 31,60 | 6,00 |
| 4 | . | . | . | . | . | . | . | . | . |
| 5 | Joven | Mujer | Primaria | Viudo | Tarde | 3,30 | 9,00 | 35,40 | 6,00 |
| 6 | Adulto | Mujer | Secun... | Viudo | Noche | 1,30 | 23,00 | 30,00 | 6,00 |
| 7 | Adulto | Mujer | Superior | Viudo | Noche | ,40 | 13,00 | 32,90 | 6,00 |
| 8 | Joven | Mujer | Primaria | Casado | Mañana | 1,50 | 12,00 | 33,20 | 6,00 |

En el ejemplo, se puede observar que el caso insertado es 4.

## 3.7. ORDENAR CASOS (DATA > SORT CASES…)




SPSS puede ordenar los casos de acuerdo a una o más variables; para ello, se ofrece una interfaz cómoda y fácil de realizar. El gráfico muestra la ventana **Sort Cases…** donde se incluye una lista de variables, a partir de la cual se puede hacer la selección. Cuando el foco azul ilumina una variable, se presionará luego el botón de desplazamiento para ubicar dicha variable en la caja **Sort by…**




El orden en que aparezcan las variables en la caja **Sort by…** determinará la jerarquía entre estas. Así, una variable será ordenada ascendente o descendentemente dentro de un valor determinado de la variable precedente.

## 3.8. DIVIDIR ARCHIVO (DATA > SPLIT FILE…)

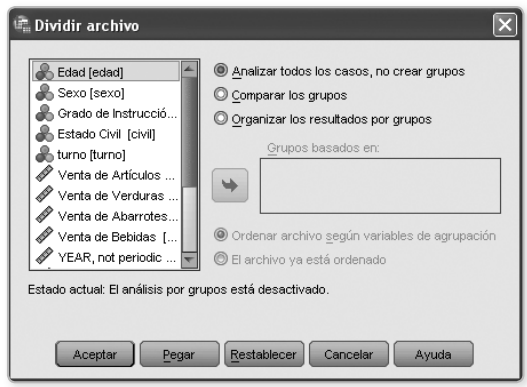
Se pueden dividir archivos presionando el ícono mostrado en la barra de herramientas, o a través del menú principal **Data > Split File…**




La segmentación de archivos es útil para realizar análisis de datos o gráficos con cruce de variables. El procedimiento es sencillo, solo se tiene que especificar la(s) variable(s) de segmentación seleccionándola(s) una por una desde la lista de variables, y presionando el botón de desplazamiento para ubicarlas en la caja **Groups based on…** (Grupos basados en). Por último, seleccione **Organize output by groups…** (Obtener resultados por grupos**)** la opción **Sort the file by grouping variables...** (Ordenar archivo según variables de segmentación).

Se puede especificar que la salida se organice de modo que permita hacer comparaciones entre los segmentos. En ese caso, se selecciona la opción **Compare groups**, segunda opción según el gráfico.

Para restablecer el análisis de todo el archivo, seleccione **Analize all cases, do not create groups…** (Analizar todos los casos, no segmentación)**.** Para desactivar la segmentación, seleccione la primera opción, Analize all cases, do not create groups. En sintaxis, la instrucción será la siguiente:



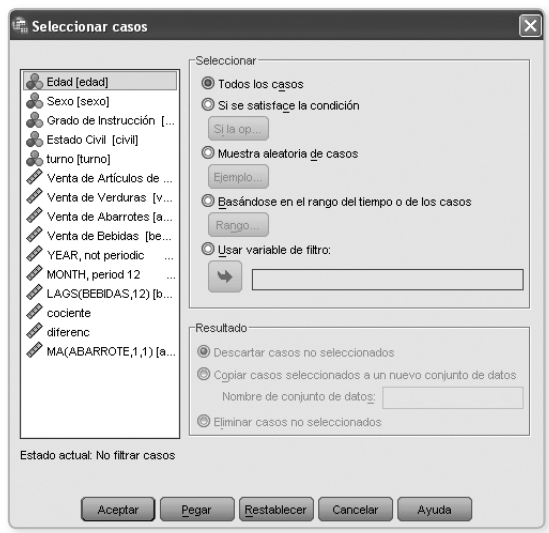

## 3.9. SELECCIONAR CASOS (DATA > SELECT CASES…)

La selección de casos de un archivo se puede realizar presionando el ícono mostrado en la barra de herramientas o a través del menú principal **Data > Select Cases…**

La **selección de datos** es una de las tareas más importantes en toda manipulación de datos. Mediante la opción **Select Cases…** SPSS le permite alcanzar este objetivo y, para ello, presenta cinco opciones:

- All cases (Todos los casos)
- If condition is satisfied (Si...)
- Random Sample of Cases (Muestra aleatoria de casos)
- Based on time or case range (Basado en rango de tiempo o de casos)
- Use filter variable (Usar variable como filtro)

En esta oportunidad, centraremos nuestra atención en la segunda y quinta opcion de ordenamiento (If condition is satisfied… y Use filter variable…, respectivamente) con los cuales son cubiertos los alcances del presente libro.



### A. If condition is satisfied…

Le permite realizar una selección de casos de acuerdo con una condición. El gráfico muestra la ventana **Seleccionar casos: Si,** donde se puede ver la lista de variables, una ventana de edición, los operadores y una lista de funciones estadísticas y de formato. La ventana **Seleccionar casos: Si** presenta un entorno gráfico que permite especificar, con facilidad y menor riesgo de cometer errores de sintaxis, la condición deseada. Por ejemplo, si se desea seleccionar los casos que presentan un salario actual mayor a 5 000 para los trabajadores de sexo femenino, la condición será la siguiente:

```
(catlab>5000) & (sexo=0)
```

Donde el símbolo "&" es el operador conjuntivo "y".

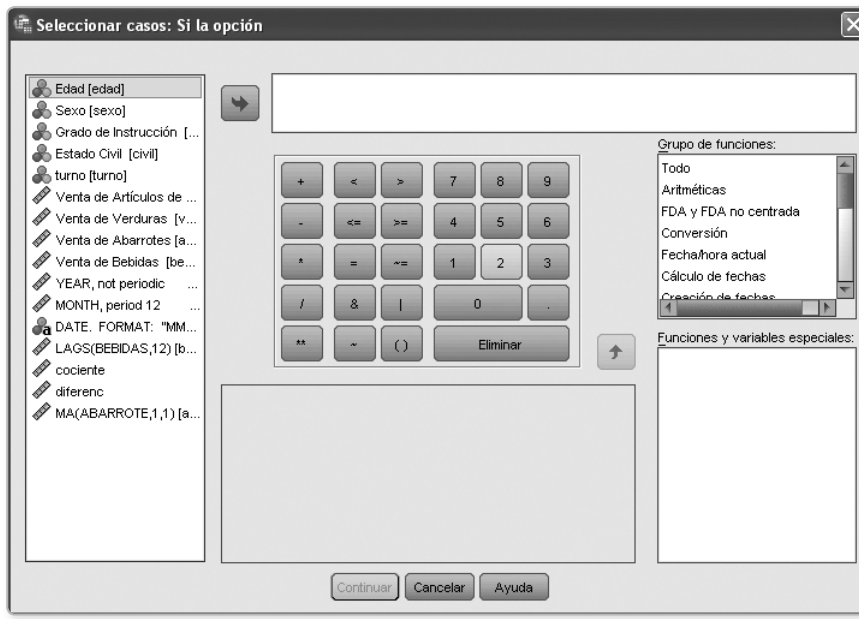



*Figura 3.1 Cuadro de diálogo Seleccionar los casos que cumplen la condición: Si...*

Además, se tiene que especificar si los casos serán filtrados o borrados. Si el caso fuera ***Filtrar*** y se regresa al Editor de datos se podrá observar que los casos no seleccionados son tachados discretamente en el índice de casos.

Si por el contrario se selecciona ***Borrar,*** los casos no seleccionados serán eliminados del Editor de datos, pudiendo así grabar en un nuevo archivo solo los casos seleccionados.

### B. Use filter variable…

También se puede realizar una selección especificando una variable de filtro. El gráfico muestra este caso en el criterio de selección que aparece al final de la lista de criterios. Si selecciona la variable Sexo de la lista de variables, el botón de desplazamiento de **Usar variable filtro** se activará. Haciendo clic en el botón de desplazamiento, dicha variable se ubicará en la respectiva línea de edición.

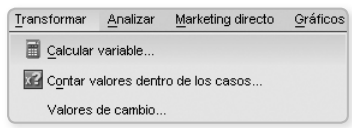
Finalmente, la selección se ejecutará presionando OK. Con el comando **Compute**, que se verá en la siguiente sección, se pueden crear filtros variables que pueden ser el producto de condiciones lógicas entre variables, y que permitirán seleccionar casos con criterios más complejos.

**NOTA**

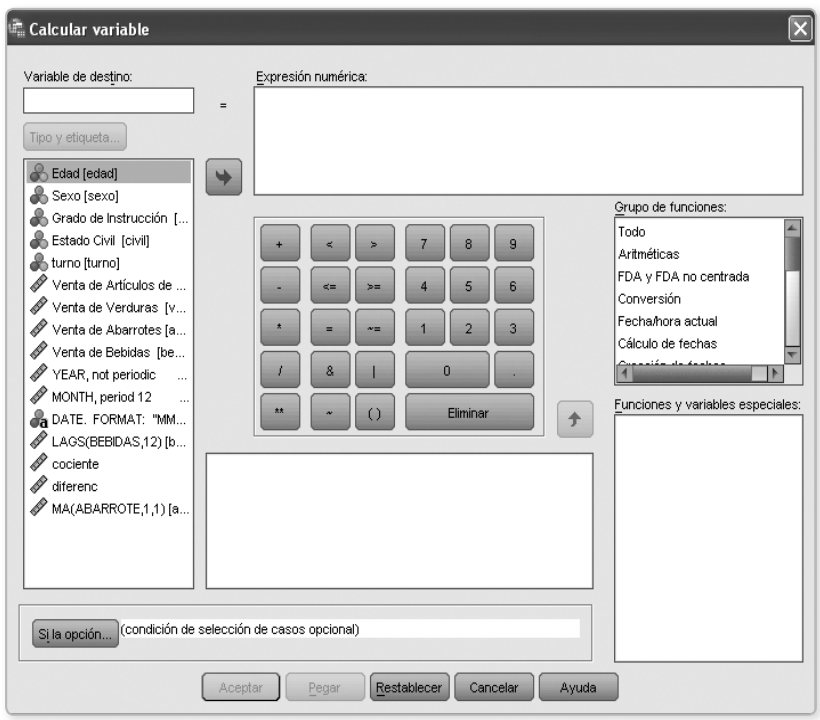
*Una explicación más detallada de cada operador puede encontrarse haciendo clic en el botón Help (Ayuda).*

## 3.10. CALCULAR VARIABLE (COMPUTE VARIABLE…)

A través del menú principal **Transfom > Compute...** se puede calcular la variable de un archivo.




Es posible crear una nueva variable que sea el producto de una transformación en la misma variable, o de otras variables existentes.

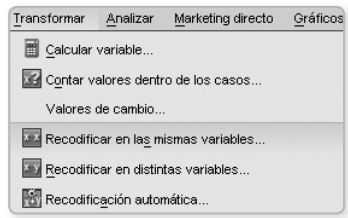



Como muestra el gráfico, la ventana **Compute variable...** está compuesta por una línea de edición para especificar el nombre de la variable objetivo, operadores (aritméticos, algebraicos, relacionales y lógicos), funciones estadísticas y matemáticas, así como la opción condicional Si (If), que permitirá restringir la transformación de variables a un subconjunto de datos que cumplen una condición dada.

Una expresión condicional puede incluir nombres de variables, constantes, operadores aritméticos, funciones estadísticas y matemáticas, y operadores lógicos y relacionales. Si se desea crear una condición que permita restringir la transformación de variables a un subconjunto de datos, se presionará el botón If. El procedimiento es el mismo que para la selección de casos visto en el acápite anterior.

## 3.11. RECODIFICAR EN LAS MISMAS VARIABLES

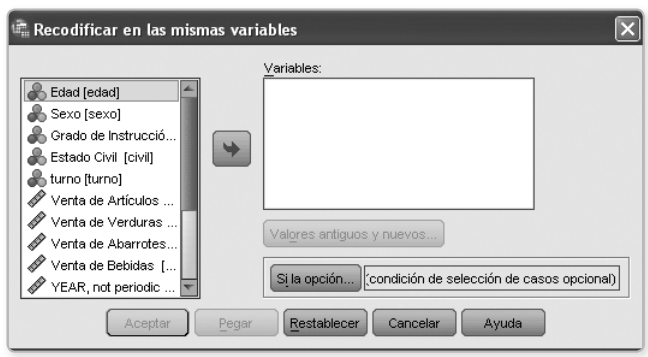
Se puede calcular la variable de un archivo a través del menú principal Transfom > Recode...



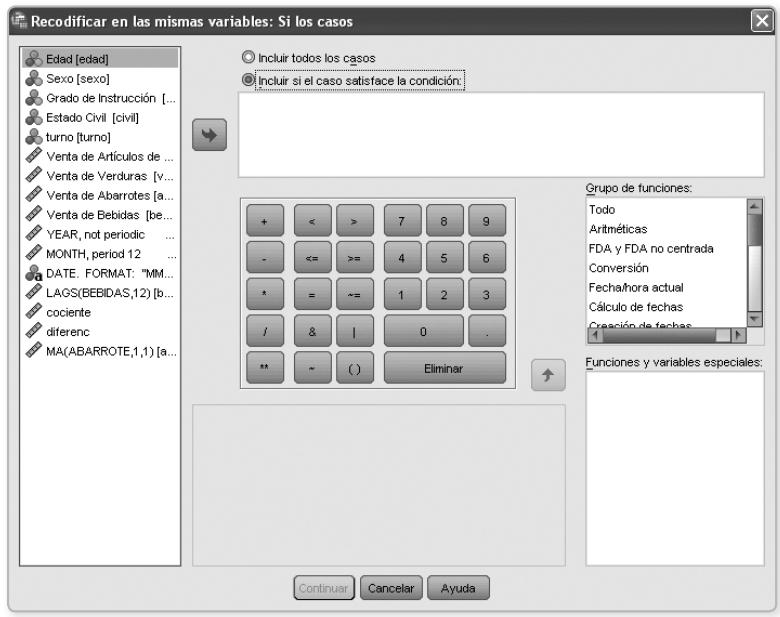

La opción (Recodificar **Recode**) transforma valores (reasigna valores) o rangos de valores de variables existentes, en nuevos valores. Estos nuevos valores o rangos de la transformación pueden asignarse en las mismas variables (Into Same Variables), o en diferentes variables (Into Different Variables).

### A. En las mismas variables (Into same variables)

En la figura siguiente se muestra la ventana del cuadro de diálogo Recodificar en las mismas variables.




Donde se puede ver la lista de variables existentes, el botón de desplazamiento de variables y la caja de edición múltiple Variables. También el cuadro de diálogo del botón de control If, cuya ventana es la siguiente:




Luego se nombra como Recode into Same Variables: If Cases y la ventana del cuadro de diálogo Old and New Variables aparece.

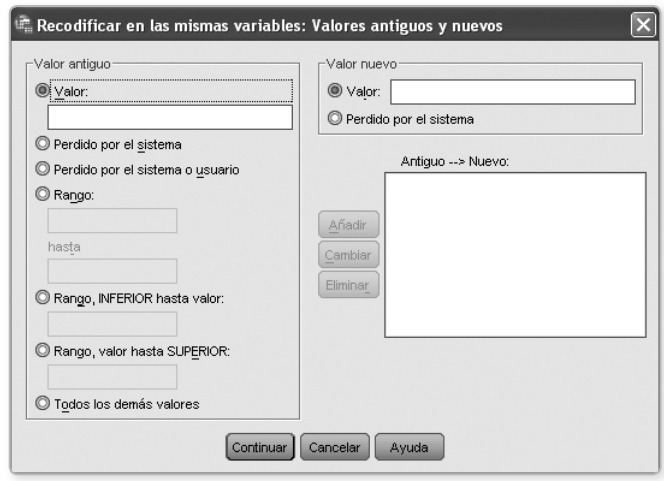



*Figura 3.2 Antiguos y nuevos valores (Old and New Values...)*

El botón de control **Old and New Values** abre la ventana mostrada en la figura anterior. Aquí la opción por defecto en la caja de verificación **Old Values** es **Value,** la cual le permite seleccionar el valor que será transformado por uno nuevo, el mismo que se especificará en la línea de edición simple de la opción **Value** de la caja de verificación **Old Value,** o por el valor perdido del sistema. Al ser completada esta información, el botón de control **Add** se activa. Presionando este, la información del antiguo y nuevo valor quedará registrada en la línea de edición múltiple **Old > New.**

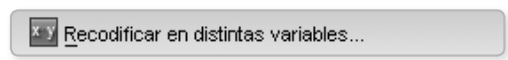
También se puede asignar un nuevo valor a los considerados perdidos por el sistema (System Missing), a los considerados perdidos por el usuario (System-or user-missing) y a un rango de valores cuyos límites superior e inferior pueden ser especificados por el usuario. Al respecto, también se puede especificar rango de valores antiguos hasta el límite superior, a partir del valor más bajo (Lowest), o especificar el límite inferior hasta el valor más alto entre los antiguos valores (highest).

Una vez especificada una serie de rangos o valores del conjunto de partida (valores antiguos) a los demás valores no considerados (*all other values*), puede reasignárseles un valor nuevo presionando la opción **All other Values** en la caja de verificación **Old-Value,** y especificando el nuevo valor en la línea de edición simple de la caja **New Value**.

Finalmente, una vez añadidos los antiguos y nuevos valores a la línea de edición multiple **Old > New** presione Aceptar y se ejecutará la transformación.

Haciendo una abstracción matemática se puede considerar como Old Values al conjunto de partida de una transformación, y a New Values como el conjunto de llegada.

## 3.12. RECODIFICAR EN DIFERENTE VARIABLE… (INTO DIFFERENT VARIABLE…)




Para llegar a esta ventana se hace clic en **Transform** > **Recode Into Different Variables**.... La siguiente figura muestra dicha ventana, que a diferencia de la ventana **Recode Into Same Variables** presenta una línea de edición múltiple **Input Variable > Output Variable;** y el botón de control **Change**, donde se ejecutará la creación de la nueva variable, cuyo nombre se puede especificar en la línea de edición simple **Output Variable**.

El botón **If** activa la misma ventana del gráfico 6 nombrado ahora como **Recode Into Different Variables: If Cases**. Del mismo modo, el botón de control **Old and New Values…** activa la misma ventana del gráfico. Presionando **OK** se ejecutará la transformación.

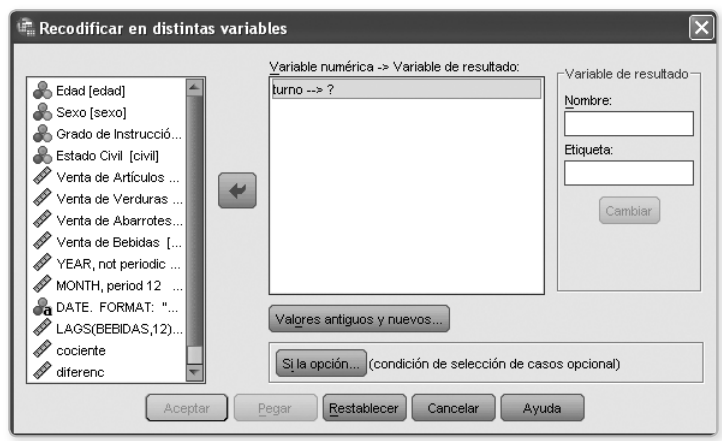



*Figura 3.3 Recodificar en Diferente variable… (Recode Into Different Variable…)*

## 3.13. RECODIFICACIÓN AUTOMÁTICA…

La decodificación automática de un archivo se puede realizar a través del menú principal **Transfom > Automatic Recode…**

En este caso, la variable a recodificar se selecciona haciendo clic sobre esta, luego se traslada a la línea de edición múltiple Variable > New Name presionando en el botón de desplazamiento. Inmediatamente se activa la línea de edición simple, donde se especificará el nombre de la nueva variable. Presionando luego el botón de control New Name, el nuevo nombre se ubicará en la línea de edición múltiple.

La recodificación puede hacerse a partir del valor más pequeño (*Lowest Value*) o del más alto (*Highest Value*), especificando la opción en la caja de verificación **Recode Starting From**. Finalmente, presionando OK, se ejecutará la transformación.

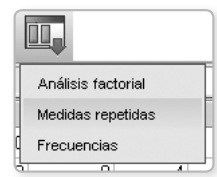



*Figura 3.4 Recodificación automática… (Automatic Recode…)*

## 3.14. VENTANA DE DIÁLOGO HERRAMIENTA MEMORIA (RECALL TOLL)

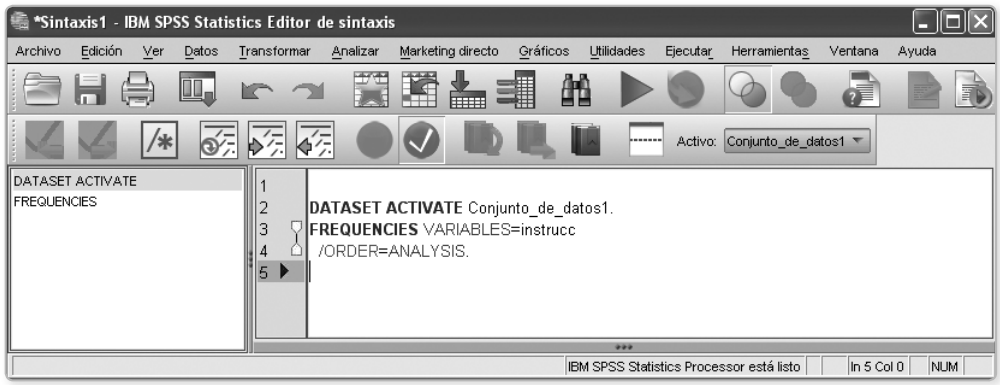


Al presionar este botón desde la barra de herramientas estándar, se abrirá una ventana donde se guardan los procedimientos utilizados. De esta forma, el usuario puede acceder de manera rápida a los procedimientos que más utiliza. El procedimiento de tablas básicas se lista como Basic Tables, y será la primera opción que se encuentre.



## 3.15. HERRAMIENTA ESTRUCTURAS DE SINTAXIS

Esta herramienta constituye una ayuda gráfica que contiene las posibles especificaciones para los diferentes comandos. Se puede acceder a ella desde la ventana de sintaxis.




## 3.16. BOTÓN EJECUTAR PROGRAMA (▶)

Al presionar este botón, se pueden ejecutar los comandos especificados en la ventana de sintaxis. Para ejecutar un comando específico, solo se necesita posicionar el cursor delante del nombre del comando.

Si por el contrario se desea ejecutar más de un comando, se tiene que seleccionar sombreando las zonas de sintaxis, arrastrando el mouse y manteniendo presionada la tecla izquierda, hasta incluir el punto final de los diferentes comandos del programa.

CAPÍTULO

4

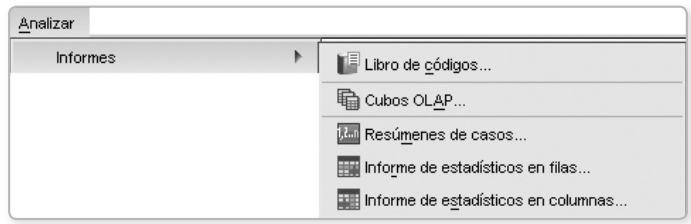
# INFORMES, ESTADÍSTICAS DESCRIPTIVAS, TABLAS

Este es un capítulo eminentemente práctico, donde se presentan distintas técnicas estadísticas como informes, estadísticas descriptivas (análisis descriptivos, distribuciones de frecuencias, medidas de tendencia central, medidas de tendencia no central, medidas de variación, gráficos de cajas y bigotes, asimetría, curtosis y tablas de contingencia) y tablas personalizadas; además, se discutirán situaciones en las que pueden aplicadas, además de las hipótesis previas necesarias para cada una de ellas y cómo realizar e interpretar los resultados.

Las potentes técnicas analíticas de SPSS le ayudarán a tomar las decisiones más adecuadas para hacer crecer su organización. La amplia gama de técnicas estadísticas permite ir más allá de las estadísticas descriptivas, así como las filas y columnas. SPSS proporciona procedimientos que permiten resolver problemas relativos a los negocios y la investigación.

## 4.1. INFORMES (REPORTS)

Informes (Reports) permite presentar con calidad los resultados obtenidos. Los **cuadros estadísticos** son un componente importante de los reportes de investigación; en consecuencia, merecen una atención especial. El presente submenú centra su atención en la presentación de información estadística mediante cuadros. Para ingresar a informes debe seguir las siguientes instrucciones (ver gráfico):




Informes presenta los procedimientos Cubos OLAP, Resúmenes de casos, Informe de estadísticos en filas e Informe de estadísticos en columnas.

### A. Cubos OLAP (procedimiento analítico interactivo)

Crea una tabla con varias capas que contienen totales, medias y otros estadísticos univariados para variables de resumen continuas, según una o más variables categóricas de agrupación. En la tabla, se creará una nueva capa para cada categoría de cada variable de agrupación.

### B. Resúmenes de casos

- Calcula estadísticos de subgrupo para las variables dentro de las categorías de una o más variables de agrupación.
- Se cruzan todos los niveles de las variables de agrupación.
- Permite elegir el orden en el que se mostrarán los estadísticos. También se muestran estadísticos de resumen para cada variable a través de todas las categorías. Los valores de los datos en cada categoría pueden mostrarse en una lista o suprimirse. Con grandes conjuntos de datos se pueden listar los primeros **n** casos.

### C. Informe de estadísticos en filas

Genera informes en los que se presentan distintos estadísticos de resumen en filas. También se encuentran disponibles listados de los casos, con o sin estadísticos de resumen.

- ***Columnas de datos:*** Son las variables del informe para obtener el listado de los casos o los estadísticos de resumen, y controlan el formato de presentación de las columnas de datos.
- ***Romper columnas por:*** Muestra una lista de las variables de ruptura opcionales que dividen el informe en grupos, y controla los estadísticos de resumen y los formatos de presentación de las columnas de ruptura.
- ***Informe.*** Controla las características globales, incluyendo los estadísticos de resumen globales, la presentación de los valores perdidos, la numeración de las páginas y los títulos.

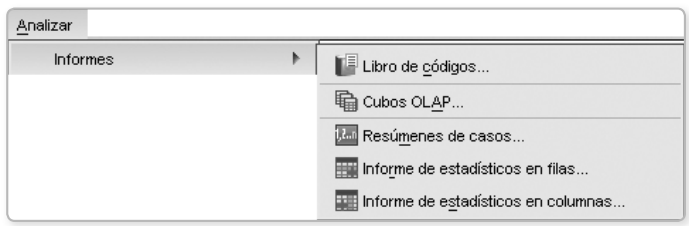
- ***Mostrar casos:*** Muestra los valores reales (o etiquetas) de las variables de la columna de datos para cada caso.
- ***Presentación preliminar:*** Muestra solo la primera página del informe. Esto es útil para examinar previamente el formato sin procesar el informe completo.
- ***Los datos ya están ordenados:*** Para los informes con variables de ruptura, los datos se deben ordenar por estas variables antes de generar el informe. Si los datos ya están ordenados, se puede ahorrar tiempo de proceso con esta opción.

**D. Informe de estadísticos en columnas**

Genera informes en los que diversos estadísticos de resumen aparecen en columnas distintas. Las opciones son similares a las del procedimiento anterior.

### 4.1.1. EJERCICIO PRÁCTICO 1

El ejercicio práctico utilizará los datos de la encuesta del supermercado *mercado.sav*, aprendiendo a realizar análisis con informes. Todos los análisis se realizan desde el menú Analizar > Informes (Reports).

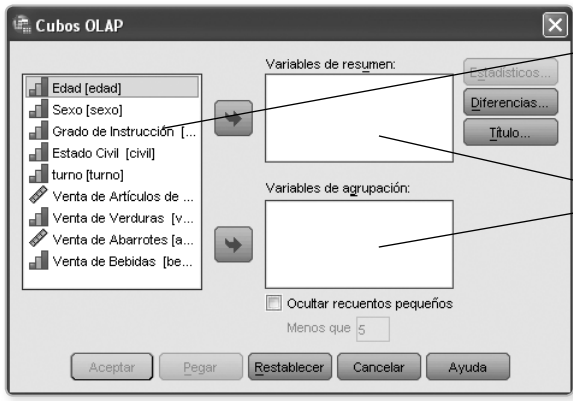



Para realizar análisis con los procedimientos de informes *(Cubos OLAP, Resúmenes de casos, Informe de estadísticas en filas e Informe de estadísticas en columnas)*, se debe seguir paso a paso los cuadros de diálogo que se presentan a continuación.

### 4.1.2. CUBOS OLAP

Las presentaciones de los informes OLAP proporcionan una forma rápida y flexible de crear, distribuir y manipular información para tomar decisiones. Usted puede crear tablas, gráficos e informes OLAP con esta herramienta, y podrá descubrir nuevos aspectos en sus datos; intercambie filas, columnas y capas de los informes OLAP o cambie rápidamente la información y las estadísticas en los gráficos, para alcanzar así nuevos niveles de comprensión. Cubos OLAP es un procedimiento que crea una tabla con varias capas que contienen totales, medias, variancias, rango y otros estadísticos.

El cuadro de diálogo Cubos OLAP ofrece el siguiente aspecto:



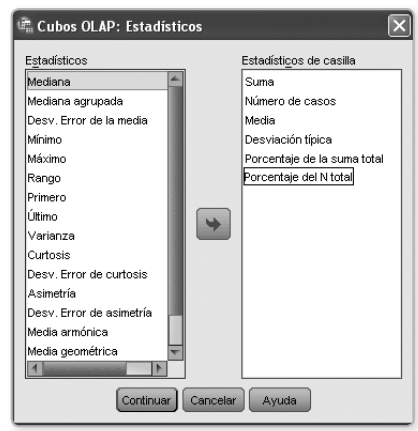

Lista de variables del archivo abierto en la ventana Editor de datos.

Trasladar aquí la variables sobre las que se quiera obtener Cubos OLAP.

En el cuadro izquierdo se muestra el listado de variables del que consta el archivo de datos que está abierto. Los cuadros de la derecha están vacíos, pues están esperando que se le indique qué variables se quiere analizar. Dichas variables se trasladan de un cuadro a otro mediante el botón . Por ejemplo, seleccione la variable que se desea analizar y luego traslade la variable a uno de los cuadros de la derecha, del modo siguiente (puede trasladarse una o más variables):




También se puede ingresar al botón Estadísticas (Statistics) y seleccionar las estadísticas que desea calcular, por ejemplo media, variancia, rango, valor mínimo, valor máximo, etc.



El cuadro de la derecha está vacío, debido a que está esperando que se le indique qué estadísticas se quiere analizar. Dichas estadísticas se trasladan de un cuadro a otro mediante el botón  . Luego pulse el botón Continuar; si desea poner títulos al cuadro de resultados, puede ingresar al procedimiento títulos de la misma forma que ingresó al procedimiento Estadísticas (Statistics). Enseguida presione el botón Aceptar (OK) para que se ejecute la orden y poder obtener el cuadro de resultados siguiente:

**OLAP Cubes**

Grado de Instrucción: Total

| | Sum | N | Mean | Std. Deviation | % of Total Sum | % of Total N |
|---|---|---|---|---|---|---|
| Venta de Artículos de Aseo Personal | 79.60 | 30 | 2.6533 | 2.2483 | 100.0% | 100.0% |
| Venta de Verduras | 376.00 | 30 | 12.5333 | 5.7219 | 100.0% | 100.0% |
| Venta de Abarrotes | 1036.20 | 30 | 34.5400 | 2.2997 | 100.0% | 100.0% |
| Venta de Bebidas | 125.00 | 30 | 4.1667 | 1.7633 | 100.0% | 100.0% |

Los resultados obtenidos aparecen en la ventana Visor de resultados. En el ejemplo se analizan la suma, media y desviación típica del Grado de Instrucción: Total, versus las Ventas de Artículos de Aseo Personal, Venta de Verduras, Venta de Abarrotes y Venta de Bebidas.

En la misma tabla se puede analizar el comportamiento del grado de instrucción de acuerdo con sus grupos, como los siguientes: No Tiene grado de instrucción, grado de instrucción Primaria, grado de instrucción Secundaria y grado de instrucción Superior. Luego haga doble clic en Grado de instrucción: Total.

**OLAP Cubes**

Grado de Instrucción: Primaria

- No Tiene
- Primaria
- Secundaria
- Superior
- Total

| | Sum | N | Mean | Std. Deviation | % of Total Sum | % of Total N |
|---|---|---|---|---|---|---|
| Venta de Artículos de Aseo Personal | 23.30 | 11 | 2.1182 | 2.1748 | 29.3% | 36.7% |
| Venta de Verduras | 120.00 | 11 | 10.9091 | 5.7871 | 31.9% | 36.7% |
| Venta de Abarrotes | 375.80 | 11 | 34.1636 | 2.0086 | 36.3% | 36.7% |
| Venta de Bebidas | 47.00 | 11 | 4.2727 | 1.4894 | 37.6% | 36.7% |

A continuación, se puede seleccionar de la lista desplegable uno de los grupos. En el caso del ejemplo, se selecciona el grupo Grado de instrucción: Primaria y el resultado se observa en la tabla siguiente:

**OLAP Cubes**

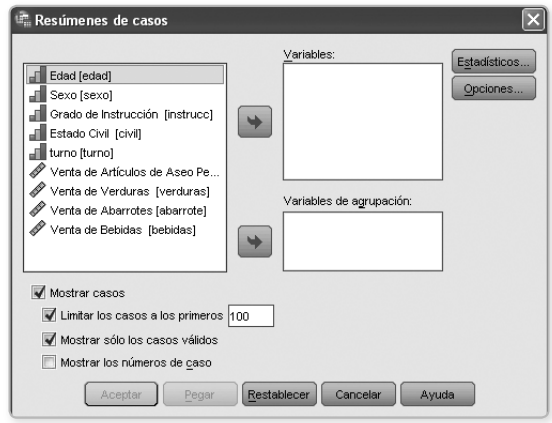
Grado de Instrucción: Primaria

| | Sum | N | Mean | Std. Deviation | % of Total Sum | % of Total N |
|---|---|---|---|---|---|---|
| Venta de Artículos de Aseo Personal | 23.30 | 11 | 2.1182 | 2.1748 | 29.3% | 36.7% |
| Venta de Verduras | 120.00 | 11 | 10.9091 | 5.7871 | 31.9% | 36.7% |
| Venta de Abarrotes | 375.80 | 11 | 34.1636 | 2.0086 | 36.3% | 36.7% |
| Venta de Bebidas | 47.00 | 11 | 4.2727 | 1.4894 | 37.6% | 36.7% |

En el ejemplo se analizan la suma, media y desviación típica, el grado de instrucción primaria vs. las Ventas de Artículos de Aseo Personal, Venta de Verduras, Venta de Abarrotes y Venta de Bebidas.

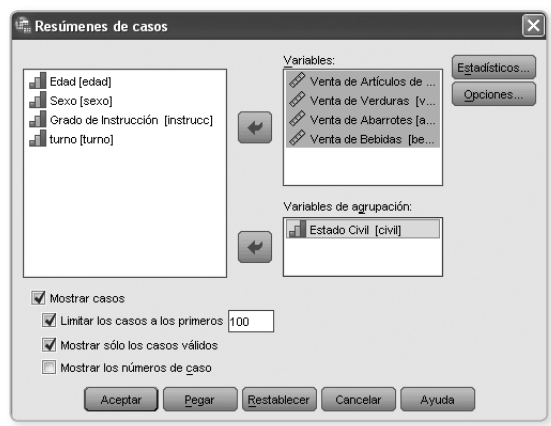
### 4.1.3. RESÚMENES DE CASOS

Calcula estadísticos de subgrupo para las variables, dentro de las categorías de una o más variables de agrupación. Permite elegir el orden en el que se mostrarán los estadísticos. Con el SPSS pueden confeccionarse Resúmenes de casos, procedimiento que crea un resumen de subgrupos y otras estadísticas. El cuadro de diálogo Resúmenes de casos ofrece el siguiente aspecto:




En el cuadro de la izquierda se muestra el listado de variables de base de datos que se está trabajando (archivo de datos abierto). Los cuadros de la derecha están vacíos, pues se está esperando que se indique qué variables se desea analizar. Dichas variables se trasladan de un cuadro a otro mediante el botón .

Observe el siguiente ejemplo: para ello, hay que trasladar las variables a los cuadros de la derecha del modo siguiente (puede trasladarse una o más variables).



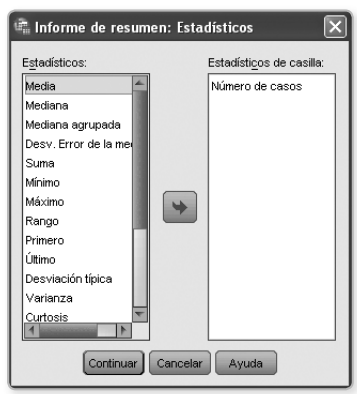

Luego pulse el botón Aceptar (OK) para que se ejecute la orden. Los resultados obtenidos aparecen en el Visor de resultados:

**Case Summaries[a]**

| | | | | Venta de Artículos de Aseo Personal | Venta de Verduras | Venta de Abarrotes |
|---|---|---|---|---|---|---|
| Estado Civil | Soltero | 1 | | 3.00 | 20.00 | 40.00 |
| | | 2 | | 1.00 | 7.00 | 34.50 |
| | | 3 | | 1.20 | 15.00 | 36.80 |
| | | 4 | | 2.30 | 6.00 | 34.50 |
| | | 5 | | 2.30 | 10.00 | 36.80 |
| | | Total | N | 5 | 5 | 5 |
| | Casado | 1 | | 2.00 | 17.00 | 34.50 |
| | | 2 | | 1.50 | 12.00 | 33.20 |
| | | 3 | | 4.50 | 19.00 | 33.10 |
| | | 4 | | .30 | 24.00 | 33.00 |
| | | 5 | | 6.00 | 10.00 | 33.20 |
| | | 6 | | .20 | 12.00 | 35.40 |
| | | 7 | | 5.50 | 6.00 | 35.90 |
| | | 8 | | .20 | 7.00 | 36.20 |
| | | 9 | | 4.50 | 18.00 | 35.40 |
| | | 10 | | .30 | 24.00 | 31.50 |
| | | 11 | | .50 | 10.00 | 33.90 |
| | | 12 | | .20 | 4.00 | 34.90 |
| | | Total | N | 12 | 12 | 12 |
| | Divorciado | 1 | | .30 | 10.00 | 31.60 |
| | | 2 | | 2.50 | 18.00 | 35.60 |
| | | 3 | | 6.00 | 14.00 | 33.20 |
| | | 4 | | 5.50 | 7.00 | 36.20 |
| | | Total | N | 4 | 4 | 4 |
| | Viudo | 1 | | 3.30 | 9.00 | 35.40 |
| | | 2 | | 1.30 | 23.00 | 30.00 |
| | | 3 | | .40 | 13.00 | 32.90 |
| | | 4 | | 5.60 | 5.00 | 31.50 |
| | | 5 | | 6.00 | 14.00 | 36.20 |
| | | 6 | | 6.50 | 9.00 | 36.50 |
| | | 7 | | .20 | 13.00 | 38.20 |
| | | 8 | | 1.50 | 13.00 | 30.10 |
| | | 9 | | 5.00 | 7.00 | 36.00 |
| | | Total | N | 9 | 9 | 9 |
| | Total | N | | 30 | 30 | 30 |

a. Limited to first 100 cases.

En el ejemplo se analiza el estado civil vs. las ventas de artículos de aseo, verduras, abarrotes y bebidas. Si se pulsa el botón Estadísticas, se obtiene el cuadro de diálogo siguiente:



El cuadro de la derecha está vacío esperando que se le indique qué estadísticas se desean analizar. Dichas estadísticas se trasladan de un cuadro a otro mediante el botón . Por ejemplo, si se traslada al cuadro de la derecha las estadísticas media, suma, mínimo, máximo y rango, se obtiene la tabla de resultados siguiente:

**Case Summaries[a]**

| | | | | Venta de Artículos de Aseo Personal | Venta de Verduras | Venta de Abarrotes |
|---|---|---|---|---|---|---|
| Estado Civil | Soltero | 1 | | 3.00 | 20.00 | 40.00 |
| | | 2 | | 1.00 | 7.00 | 34.50 |
| | | 3 | | 1.20 | 15.00 | 36.80 |
| | | 4 | | 2.30 | 6.00 | 34.50 |
| | | 5 | | 2.30 | 10.00 | 36.80 |
| | | Total | N | 5 | 5 | 5 |
| | | | Median | 2.3000 | 10.0000 | 36.8000 |
| | | | Sum | 9.80 | 58.00 | 182.60 |
| | | | Minimum | 1.00 | 6.00 | 34.50 |
| | | | Maximum | 3.00 | 20.00 | 40.00 |
| | | | Range | 2.00 | 14.00 | 5.50 |
| | Casado | 1 | | 2.00 | 17.00 | 34.50 |
| | | 2 | | 1.50 | 12.00 | 33.20 |
| | | 3 | | 4.50 | 19.00 | 33.10 |
| | | 4 | | .30 | 24.00 | 33.00 |
| | | 5 | | 6.00 | 10.00 | 33.20 |
| | | 6 | | .20 | 12.00 | 35.40 |
| | | 7 | | 5.50 | 6.00 | 35.90 |
| | | 8 | | .20 | 7.00 | 36.20 |
| | | 9 | | 4.50 | 18.00 | 35.40 |
| | | 10 | | .30 | 24.00 | 31.50 |
| | | 11 | | .50 | 10.00 | 33.90 |
| | | 12 | | .20 | 4.00 | 34.90 |
| | | Total | N | 12 | 12 | 12 |
| | | | Median | 1.0000 | 12.0000 | 34.2000 |
| | | | Sum | 25.70 | 163.00 | 410.20 |
| | | | Minimum | .20 | 4.00 | 31.50 |
| | | | Maximum | 6.00 | 24.00 | 36.20 |
| | | | Range | 5.80 | 20.00 | 4.70 |
| | Divorciado | 1 | | .30 | 10.00 | 31.60 |
| | | 2 | | 2.50 | 18.00 | 35.60 |
| | | 3 | | 6.00 | 14.00 | 33.20 |
| | | 4 | | 5.50 | 7.00 | 36.20 |
| | | Total | N | 4 | 4 | 4 |
| | | | Median | 4.0000 | 12.0000 | 34.4000 |
| | | | Sum | 14.30 | 49.00 | 136.60 |
| | | | Minimum | .30 | 7.00 | 31.60 |
| | | | Maximum | 6.00 | 18.00 | 36.20 |
| | | | Range | 5.70 | 11.00 | 4.60 |
| | Viudo | 1 | | 3.30 | 9.00 | 35.40 |
| | | 2 | | 1.30 | 23.00 | 30.00 |
| | | 3 | | .40 | 13.00 | 32.90 |
| | | 4 | | 5.60 | 5.00 | 31.50 |
| | | 5 | | 6.00 | 14.00 | 36.20 |
| | | 6 | | 6.50 | 9.00 | 36.50 |
| | | 7 | | .20 | 13.00 | 38.20 |
| | | 8 | | 1.50 | 13.00 | 30.10 |
| | | 9 | | 5.00 | 7.00 | 36.00 |
| | | Total | N | 9 | 9 | 9 |
| | | | Median | 3.3000 | 13.0000 | 35.4000 |
| | | | Sum | 29.80 | 106.00 | 306.80 |
| | | | Minimum | .20 | 5.00 | 30.00 |
| | | | Maximum | 6.50 | 23.00 | 38.20 |
| | | | Range | 6.30 | 18.00 | 8.20 |
| | Total | N | | 30 | 30 | 30 |
| | | Median | | 2.1500 | 12.0000 | 34.7000 |
| | | Sum | | 79.60 | 376.00 | 1036.20 |
| | | Minimum | | .20 | 4.00 | 30.00 |
| | | Maximum | | 6.50 | 24.00 | 40.00 |
| | | Range | | 6.30 | 20.00 | 10.00 |

a. Limited to first 100 cases.

## 4.2. ESTADÍSTICOS DESCRIPTIVOS

Es la herramienta estadística más utilizada en diferentes niveles básicos de estudio. Estadísticas descriptivas en el SPSS presenta las siguientes herramientas: Frecuencia, Descriptivos, Explorar, Tablas de contingencia, Razón, Gráficos P-P y Gráficos Q-Q.




### A. Frecuencias

Presenta estadísticas y representaciones gráficas útiles para describir o inspeccionar muchos tipos de variables. Los valores pueden organizarse en orden ascendente o descendente, y se pueden ordenar las categorías por sus frecuencias. Se puede suprimir el informe de frecuencias cuando una variable posee muchos valores diferentes y permite etiquetar los gráficos con las frecuencias o los porcentajes.

### B. Descriptivos

Proporciona estadísticas de resumen para varias variables en una única tabla, y calcula valores tipificados (puntuaciones Z). Las variables se pueden ordenar por el tamaño de sus medias (en orden ascendente o descendente), alfabéticamente o por el orden en el que se seleccionen (por defecto). Se pueden guardar las puntuaciones Z, añadiéndose al Editor de datos para su posterior uso.

### C. Explorar

Genera estadísticas de resumen y representaciones gráficas, bien para todos los casos o bien de forma separada para grupos de casos. Este procedimiento se puede usar para inspeccionar los datos, identificar valores atípicos o extremos, descubrir discontinuidades, obtener descripciones, comprobar supuestos y caracterizar diferencias entre subpoblaciones (grupos de casos). La exploración de los datos puede ayudar a determinar si las técnicas de análisis que se están empleando son adecuadas.

### D. Tablas de contingencia

Crea tablas de clasificación doble y múltiple, y además proporciona una serie de pruebas y medidas de asociación para las tablas de doble clasificación. La estructura de la tabla y el hecho de que las categorías estén ordenadas o no, determinan las pruebas o medidas que se utilizan. Las tablas de contingencia y las medidas de asociación solo se calculan para las tablas de doble clasificación; si especifica una fila, columna y factor de capa (variable de control), se creará un panel de medidas y estadísticos asociados para cada valor del factor de capa (o una combinación de valores para dos o más variables de control).

### E. Razón (Ratios)

Permite calcular *ratios estadísticos* como los coeficientes de dispersión, coeficientes de variación y otros, permitiendo facilitar enormemente a los usuarios la interpretación de indicadores estadísticos que pueden servir para una toma de decisiones rápida y oportuna.

### F. Gráficos P-P (Plots P-P)

Crea un gráfico de las proporciones acumuladas de una variable respecto a las de una distribución cualquiera de prueba; es decir, por cada variable especificada proporciona dos gráficos: en el primero se representan los valores de la función de distribución acumulativa esperada, bajo el supuesto de normalidad (por defecto) frente a los observados. En el segundo, se representan los residuos. Los gráficos de probabilidad suelen emplearse para determinar si la distribución de una variable coincide con una distribución dada.

### G. Gráficos Q-Q (Plots Q-Q)

Crea un gráfico con los cuantiles de distribución de una variable, respecto a los cuantiles de una distribución cualquiera de prueba; es decir, por cada variable especificada proporciona dos gráficos: en el primero, se representan los valores de los cuantiles esperados, bajo el supuesto de distribución normal (por defecto) de media cero y varianza uno frente a los observados.

## 4.2.1. EJERCICIO PRÁCTICO 1

En el presente ejercicio se utilizan los datos de la encuesta Supermercado, mostrada en el capitulo anterior *mercado.sav*.

### A. Análisis estadísticos descriptivos

Se aprende a realizar análisis descriptivos con una variable. Como ya se vio anteriormente, todos los procesos de análisis se realizan desde el menú Analizar > Estadísticas descriptivas.




Si solo se quiere realizar análisis descriptivos univariados, bastará con los tres primeros procedimientos del SPSS (Frecuencias, Descriptivos y Explorar).

Antes de continuar, es necesario definir algunos conceptos importantes.

### B. Estadística descriptiva

Es una ciencia que analiza series de datos (por ejemplo: edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) y trata de extraer conclusiones sobre el comportamiento de estas variables.

Por otro lado, las variables pueden ser de dos tipos:

- **Variables cualitativas o atributos**. No se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo, etc.).
- **Variables cuantitativas**. Tienen valor numérico (edad, precio de un producto, ingresos anuales). Las variables cuantitativas se pueden clasificar en *discretas y continuas*.
- **Discretas.** Solo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3...; pero nunca podrá ser 3.45, por ejemplo).
  - **Continuas**. Pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h...etc.

Las variables también se pueden clasificar en:

- **Variables unidimensionales.** Solo recogen información sobre una característica (por ejemplo: edad de los alumnos de una clase).
- **Variables bidimensionales.** Recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).
- **Variables pluridimensionales.** Recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

### C. Frecuencias (distribución de frecuencia)

#### a. Distribución de frecuencia

Es la representación estructurada en forma de tabla, de toda la información que se ha recogido sobre la variable que se desea estudiar, de la siguiente manera:

| Variable | Frecuencias absolutas | | Frecuencias relativas | |
|---|---|---|---|---|
| (Valores) | Simple | Acumulada | Simple | Acumulada |
| $X_1$ | $f_1$ | $f_1$ | $h_1 = f_1/n$ | $h_1$ |
| $X_2$ | $f_2$ | $f_1 + f_2$ | $h_2 = f_2/n$ | $h_1 + h_2$ |
| ... | ... | ... | ... | ... |
| $X_{k-1}$ | $f_{k-1}$ | $f_1 + f_2 + ... + f_{k-1}$ | $h_{k-1} = f_{k-1}/n$ | $h_1 + h_2 + ... + h_{k-1}$ |
| $X_k$ | $f_k$ | $\Sigma f_i = n$ | $h_k = f_k/n$ | 1 |

Donde:

$X_i$ : Son los distintos valores que puede tomar la variable. (i = 1, 2, 3... k).
$f_i$ : Es el número de veces que se repite cada valor.
$h_i$ : Es el porcentaje que la repetición de cada valor supone sobre el total.
n : Es el número de observaciones realizadas (tamaño de la muesta).

**Ejemplo:**

Suponga que se mide la altura de los niños de una clase y se obtienen los siguientes resultados (cm):

| Alumno | Estatura | Alumno | Estatura | Alumno | Estatura |
|---|---|---|---|---|---|
| Alumno 1 | 1,25 | Alumno 11 | 1,23 | Alumno 21 | 1,21 |
| Alumno 2 | 1,28 | Alumno 12 | 1,26 | Alumno 22 | 1,29 |
| Alumno 3 | 1,27 | Alumno 13 | 1,30 | Alumno 23 | 1,26 |
| Alumno 4 | 1,21 | Alumno 14 | 1,21 | Alumno 24 | 1,22 |
| Alumno 5 | 1,22 | Alumno 15 | 1,28 | Alumno 25 | 1,28 |
| Alumno 6 | 1,29 | Alumno 16 | 1,30 | Alumno 26 | 1,27 |
| Alumno 7 | 1,30 | Alumno 17 | 1,22 | Alumno 27 | 1,26 |
| Alumno 8 | 1,24 | Alumno 18 | 1,25 | Alumno 28 | 1,23 |
| Alumno 9 | 1,27 | Alumno 19 | 1,20 | Alumno 29 | 1,22 |
| Alumno 10 | 1,29 | Alumno 20 | 1,28 | Alumno 30 | 1,21 |

Si presenta esta información estructurada en una tabla de frecuencias, se obtendrá la siguiente tabla:

**Tabla de frecuencias**

| Variable | Frecuencias absolutas | | Frecuencias relativas | |
|---|---|---|---|---|
| (Valor) | Simple | Acumulada | Simple | Acumulada |
| 1,20 | 1 | 1 | 3,3% | 3,3% |
| 1,21 | 4 | 5 | 13,3% | 16,6% |
| 1,22 | 4 | 9 | 13,3% | 30,0% |
| 1,23 | 2 | 11 | 6,6% | 36,6% |
| 1,24 | 1 | 12 | 3,3% | 40,0% |
| 1,25 | 2 | 14 | 6,6% | 46,6% |
| 1,26 | 3 | 17 | 10,0% | 56,6% |
| 1,27 | 3 | 20 | 10,0% | 66,6% |
| 1,28 | 4 | 24 | 13,3% | 80,0% |
| 1,29 | 3 | 27 | 10,0% | 90,0% |
| 1,30 | 3 | 30 | 10,0% | 100,0% |

Si los valores que toma la variable son muy diversos y cada uno de ellos se repite muy pocas veces, entonces conviene agruparlos por intervalos, ya que de otra manera, se obtendría una tabla de frecuencia muy extensa, que aportaría muy poco valor para efectos de síntesis de la información.

Para obtener los resultados con el SPSS, vaya al menú Analizar, luego Estadísticos descriptivos y haga clic en Frecuencias. Se obtiene la tabla siguiente:

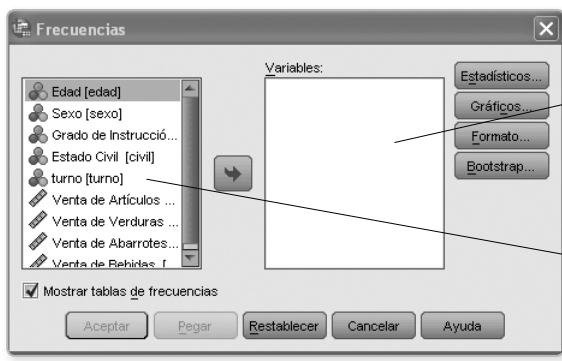
**Estatura**

| | | Frecuencia | Porcentaje | Porcentaje válido | Porcentaje acumulado |
|---|---|---|---|---|---|
| Válidos | 1,20 | 1 | 3,3 | 3,3 | 3,3 |
| | 1,21 | 4 | 13,3 | 13,3 | 16,7 |
| | 1,22 | 4 | 13,3 | 13,3 | 30,0 |
| | 1,23 | 2 | 6,7 | 6,7 | 36,7 |
| | 1,24 | 1 | 3,3 | 3,3 | 40,0 |
| | 1,25 | 2 | 6,7 | 6,7 | 46,7 |
| | 1,26 | 3 | 10,0 | 10,0 | 56,7 |
| | 1,27 | 3 | 10,0 | 10,0 | 66,7 |
| | 1,28 | 4 | 13,3 | 13,3 | 80,0 |
| | 1,29 | 3 | 10,0 | 10,0 | 90,0 |
| | 1,30 | 3 | 10,0 | 10,0 | 100,0 |
| | Total | 30 | 100,0 | 100,0 | |

A continuación, se detalla cómo se obtienen los resultados de tabla de frecuencias con SPSS.

### D. DISTRIBUCIÓN DE FRECUENCIA CON SPSS

SPSS puede confeccionar distribuciones de frecuencias para organizar los datos correspondientes a una variable. En este apartado, se verá cómo se construyen las tablas de frecuencias y cómo se realizan algunas de las representaciones gráficas. El cuadro de diálogo **Frecuencias** ofrece el siguiente aspecto:

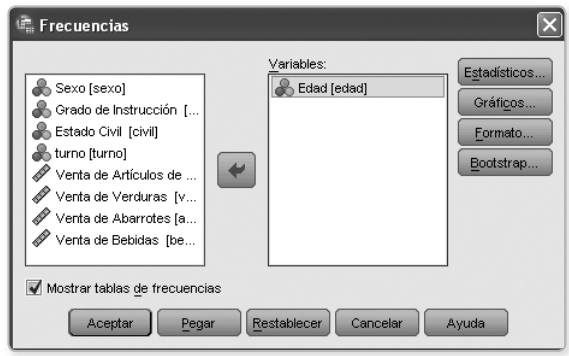



Trasladar aquí las variables sobre las que se quiere obtener la distribución de frecuencias.

Lista de variables del archivo abierto en la ventana Editor de datos.

El cuadro de la izquierda muestra el listado de las variables de la base de datos en estudio. El cuadro derecho está vacío, esperando la indicación de las variables a analizar. Las variables se trasladan de un cuadro a otro mediante el botón . Por defecto, la opción Mostrar tablas de frecuencias (Display frequency tables) está activada; sin embargo, si no desea que aparezca, puede desactivarla.

Para trabajar con la variable Edad, se traslada la variable al cuadro de la derecha del modo siguiente:




Luego, pulse el botón Aceptar (OK) para que se ejecute la orden. Los resultados obtenidos aparecen en el Visor de resultados. La tabla de frecuencias tiene el siguiente aspecto:

**Edad**

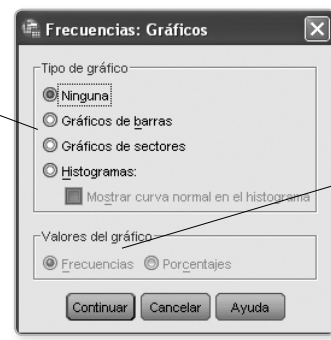
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Adolescente | 3 | 10.0 | 10.0 | 10.0 |
| | Joven | 5 | 16.7 | 16.7 | 26.7 |
| | Adulto | 22 | 73.3 | 73.3 | 100.0 |
| | Total | 30 | 100.0 | 100.0 | |

En la primera columna, aparecen los valores de la variable (Adolescente, Joven y Adulto); en la segunda columna, las frecuencias absolutas en sentido ascendente (o descendente); en la tercera columna, las frecuencias relativas en porcentaje; y, en la última columna (porcentaje acumulado), las frecuencias relativas acumuladas en porcentaje.

### a. Representaciones gráficas

También es posible pedir al SPSS que realice tres tipos de gráficos para observar las frecuencias de los datos. Desde el cuadro de diálogo anterior, presione Gráficos (Charts...), obteniendo el siguiente cuadro de diálogo:

Solo puede seleccionarse uno; luego pulse en Continuar.

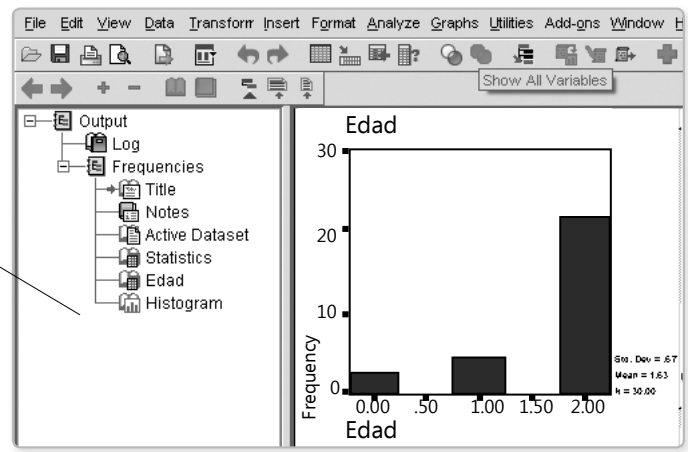



Permite confeccionar el gráfico con las frecuencias absolutas o relativas.

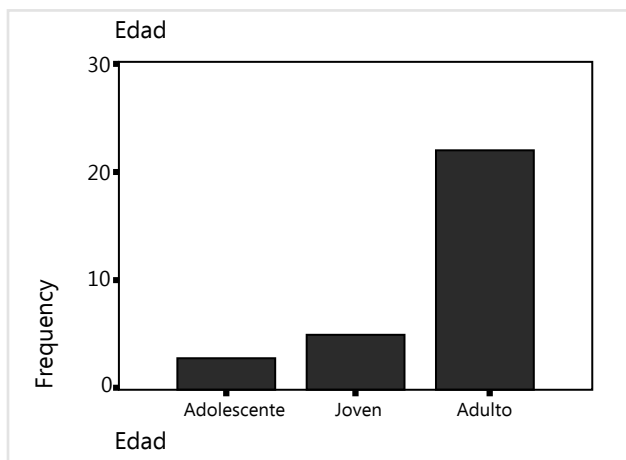
Como se observa, se puede obtener un histograma (con el ajuste a la curva normal), un diagrama de barras (Bar charts) y un diagrama de sectores o pictograma (Pie chats). Desde este cuadro de diálogo, solo es posible seleccionar uno de los gráficos. Si se quisiese obtener los tres, hay que repetir la operación tres veces.

Continuando con la variable Edad, ahora se va a solicitar a SPSS que muestre los tres tipos de gráficos de la distribución de frecuencias para la edad. Si pulsa Continuar (OK), se vuelve automáticamente al cuadro de diálogo anterior. Para ir de un resultado a otro, puede utilizarse la parte izquierda de la pantalla, donde aparece el listado de resultados obtenidos (en el ejemplo: Edad e Histograma). Situándose con el mouse en cada uno de ellos, puede accederse automáticamente al resultado correspondiente. La siguiente figura muestra el histograma para la variable Edad:

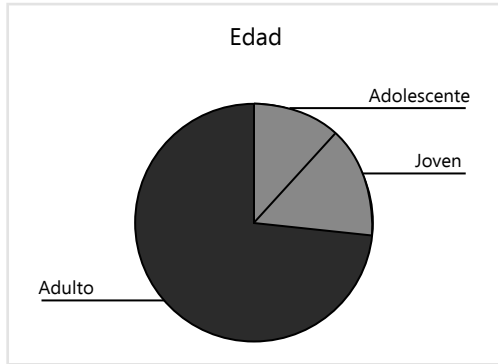
Para ir de un resultado a otro (por ejemplo, a la tabla de frecuencias o al histograma), sitúese con el mouse en el título correspondiente (en el ejemplo, Histograma).




Se puede pedir al SPSS graficar el diagrama de barras para la variable Edad.



También es posible graficar el diagrama de sectores para la variable Edad. El resultado obtenido es el siguiente:




En el gráfico de sectores, se observa que los adultos son los que más compran en el supermercado, seguidos por los jóvenes. Como ejercicio, usted puede realizar gráficos (histograma, barras, sectores), teniendo presente que el gráfico debe ser el que mejor describa la variable que se está analizando.

### E. Medidas de posición central

Informan sobre los valores medios de la serie de datos. Las principales medidas de posición central son las siguientes:

#### a. Media

Es el valor medio ponderado de la serie de datos. Se pueden calcular diversos tipos de media, siendo las más utilizadas:

- **Media aritmética.** Se calcula multiplicando cada valor por el número de veces que se repite. La suma de todos estos productos se divide por el total de datos de la muestra.

$$\overline{X}a = \frac{(X1*f1)+(X2*f2)+(X3*f3)+\ldots+(Xn-1*fn-1)+(Xn*fn)}{n}$$

- **Media geométrica.** Se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto final se le calcula la raíz "n" (siendo "n" el total de datos de la muestra).

$$\overline{X}m = \left( X_1^{f_1} * X_2^{f_2} * X_3^{f_3} * \ldots\ldots * X_k^{f_{k1}} \right)^{(1/n)}$$

Según el tipo de datos que se analice, será más apropiado utilizar la media aritmética o la media geométrica. La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la **media aritmética** es la medida de posición central más utilizada. Lo más positivo de la media es que, en su cálculo, se utilizan todos los valores de la serie. por lo que no se pierde alguna información.

**b. Mediana**

Es el valor de la serie de datos que se sitúa precisamente en el centro de la muestra (un 50% de valores son inferiores y otro 50% son superiores).

**c. Moda**

Es el valor que se repite con mayor frecuencia en la muestra.

## Ejemplo:

Se va a utilizar la tabla de distribución de frecuencias con los datos de la estatura de los alumnos, tomados del ejemplo anterior.

| Variable | Frecuencias absolutas | | Frecuencias relativas | |
|---|---|---|---|---|
| (Valor) | Simple | Acumulada | Simple | Acumulada |
| 1,20 | 1 | 1 | 3,3% | 3,3% |
| 1,21 | 4 | 5 | 13,3% | 16,6% |
| 1,22 | 4 | 9 | 13,3% | 30,0% |
| 1,23 | 2 | 11 | 6,6% | 36,6% |
| 1,24 | 1 | 12 | 3,3% | 40,0% |
| 1,25 | 2 | 14 | 6,6% | 46,6% |
| 1,26 | 3 | 17 | 10,0% | 56,6% |
| 1,27 | 3 | 20 | 10,0% | 66,6% |
| 1,28 | 4 | 24 | 13,3% | 80,0% |
| 1,29 | 3 | 27 | 10,0% | 90,0% |
| 1,30 | 3 | 30 | 10,0% | 100,0% |

Calcule los valores de las distintas medidas de posiciones centrales:

- **Media aritmética**. 1.253. Es decir, la estatura media de este grupo de alumnos es 1,253 cm.
- **Media geométrica**. 1.253. En este ejemplo, la media aritmética y la media geométrica coinciden, aunque no siempre esto es así.
- **Mediana**. La mediana de esta muestra es 1,26 cm, ya que por debajo está el 50% de los valores y, por arriba, el otro 50%. Esto se puede apreciar al analizar la columna de frecuencias relativas acumuladas. En este ejemplo, como el valor 1,26 se repite en tres ocasiones, la media se situará exactamente entre el primer y el segundo valor de este grupo, debido a que entre estos dos valores, se encuentra la división entre el 50% inferior y el 50% superior.
- **Moda**. Hay tres valores que se repiten en cuatro ocasiones: el 1,21, el 1,22 y el 1,28; por lo tanto, esta sería cuenta con tres modas.

Resultados con el SPSS:

**Estadísticos**

Estatura

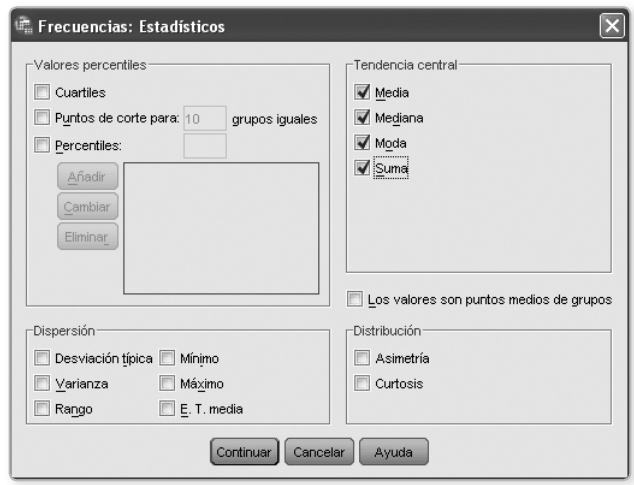
| | | |
|---|---|---|
| N | Válidos | 30 |
| | Perdidos | 0 |
| Media | | 1,2533 |
| Mediana | | 1,2600 |
| Moda | | 1,21[a] |
| Suma | | 37,60 |

a. Existen varias modas. Se mostrará el menor de los valores.

### F. Medidas de posición central con SPSS (media, mediana, moda y suma)

Estos estadísticos pueden obtenerse prácticamente desde cualquier cuadro de diálogo del menú **Analizar > Estadísticas descriptivas**. Uno de los más empleados es el del cuadro de diálogo Frecuencias, ya que cuando se elabora una tabla de frecuencias, normalmente también se desea obtener las medidas de posición.

Para ello, se pulsa el botón Estadísticas, luego aparece el cuadro de diálogo. Aquí se seleccionan las estadísticas que se desean utilizar en el estudio. En el ejemplo dado, se pide calcular la media, mediana, moda y suma. Esto se logra simplemente marcando cada opción.




Luego se presiona el botón Continuar y después Aceptar (OK). El SPSS calcula estos cuatro valores estadísticos para el listado de variables seleccionadas.

**Statistics**

Venta de Artículos de Aseo Persona

| | | |
|---|---|---|
| N | Valid | 30 |
| | Missing | 0 |
| Mean | | 2.6533 |
| Median | | 2.1500 |
| Mode | | .20 |
| Sum | | 79.60 |

La media que se obtiene se corresponde con la media aritmética; la mediana, con el valor central cuando los datos se ordenan en modo creciente; y la moda, con el valor que obtiene la frecuencia absoluta mayor.

Para conocer las estadísticas descriptivas de las variables también puede seleccionarse el procedimiento **Descriptivas**. Este procedimiento presenta un cuadro de diálogo parecido al del procedimiento Frecuencias. El cuadro de diálogo del procedimiento **Descriptivas** es el siguiente:




Pulse este botón para indicar qué estadísticos univariados se desea obtener.

Para obtener las estadísticas descriptivas, traslade las variables correspondientes al cuadro Variables y luego pulse Opciones, obteniéndose el cuadro de diálogo siguiente:




En este caso, solamente está disponible la estadística de tendencia central media y suma que, como antes, representa la media aritmética para las variables seleccionadas. Observe el siguiente ejemplo para algunas variables del archivo:

**Descriptive Statistics**

| | N | Sum | Mean |
|---|---|---|---|
| Venta de Artículos de Aseo Personal | 30 | 79.60 | 2.6533 |
| Venta de Verduras | 30 | 376.00 | 12.5333 |
| Venta de Abarrotes | 30 | 1036.20 | 34.5400 |
| Venta de Bebidas | 30 | 125.00 | 4.1667 |
| Valid N (listwise) | 30 | | |

Se aprecia que las 30 personas gastaron en total *S/.*79.60 en artículos de aseo personal, en el supermercado la media de los gastos es S/.2.6533. En forma similar, se aprecia en la venta de verduras que las 30 personas gastaron *S/.*356.00 en total, con una media o promedio de gastos de S/.12.533.

### G. Medidas de posición no central

Informan sobre cómo se distribuye el resto de los valores de la serie. Las medidas de posición no central permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre otros indicadores, se suele utilizar una serie de valores que dividen la muestra en tramos iguales, como:

- **Cuartiles.** Son tres valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente.
- **Deciles.** Son nueve valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.
- **Percentiles.** Son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados.

## Ejemplo:

Calcular los cuartiles de la serie de datos referidos a la estatura de un grupo de alumnos (lección 2.ª). Los deciles y centiles se calculan de igual manera, aunque harán falta distribuciones con mayor número de datos.

| Variable | Frecuencias absolutas | | Frecuencias relativas | |
|---|---|---|---|---|
| (Valor) | Simple | Acumulada | Simple | Acumulada |
| 1,20 | 1 | 1 | 3,3% | 3,3% |
| 1,21 | 4 | 5 | 13,3% | 16,6% |
| 1,22 | 4 | 9 | 13,3% | 30,0% |
| 1,23 | 2 | 11 | 6,6% | 36,6% |
| 1,24 | 1 | 12 | 3,3% | 40,0% |
| 1,25 | 2 | 14 | 6,6% | 46,6% |
| 1,26 | 3 | 17 | 10,0% | 56,6% |
| 1,27 | 3 | 20 | 10,0% | 66,6% |
| 1,28 | 4 | 24 | 13,3% | 80,0% |
| 1,29 | 3 | 27 | 10,0% | 90,0% |
| 1,30 | 3 | 30 | 10,0% | 100,0% |

## Solución:

- **1.er cuartil**. Es el valor 1,22 cm, ya que por debajo de ella se sitúa el 25% de la frecuencia (tal como se puede ver en la columna de la frecuencia relativa acumulada).
- **2.do cuartil**. Es el valor 1,26 cm, ya que, entre este valor y el primer cuartil, se sitúa otro 25% de la frecuencia.
- **3.er cuartil**. Es el valor 1,28 cm, dado que entre este valor y el segundo cuartil, se sitúa otro 25% de la frecuencia. Además, por encima de él queda el restante 25% de la frecuencia.

**NOTA**

*Cuando un cuartil recae en un valor que se ha repetido más de una vez (como ocurre en el ejemplo en los tres cuartiles), la medida de posición no central será realmente una de las repeticiones.*

Resultados con el SPSS:

**Estadísticos**

Estatura

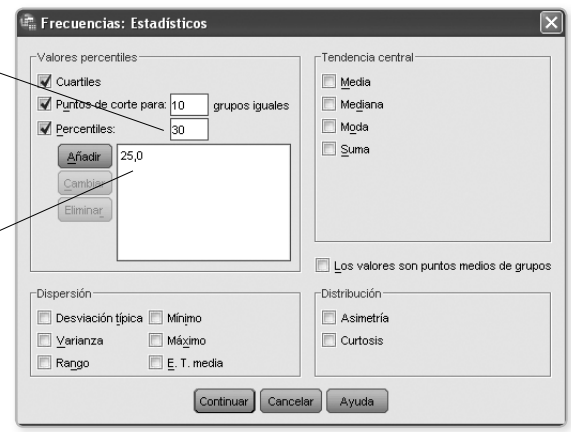
| | | |
|---|---|---|
| N | Válidos | 30 |
| | Perdidos | 0 |
| Percentiles | 25 | 1,2200 |
| | 50 | 1,2600 |
| | 75 | 1,2800 |

### H. Medidas de posición no central con SPSS

Comprende los percentiles (centiles), los deciles y los cuartiles. Los centiles y otros cuantiles pueden obtenerse desde el procedimiento Frecuencias. En la parte inferior de este cuadro de diálogo hay un botón llamado Estadísticos, si se selecciona se obtiene el siguiente cuadro de diálogo:

Si se desea obtener los cuartiles, indique en grupos de 4; para los deciles, 10; y para los centiles, 100.

Esta opción es específica para los centiles y permite definir el centil (entre 1 y 99) que se desea obtener. Para indicar varios, escriba el número, pulse en Añadir (Add) y así sucesivamente.




Desde aquí puede indicarse que se obtengan diferentes tipos de cuantiles o percentiles. Se puede indicar el número exacto del percentil que se desee obtener e ir añadiendo a la lista inferior. Continúe con la variable artículos de aseo personal. Si le pide que le dé los cuartiles, deciles y los percentiles 25 y 30, se obtiene lo siguiente:

**Statistics**

Venta de Artículos de Aseo Personal

| | | |
|---|---|---|
| N | Valid | 30 |
| | Missing | 0 |
| Percentiles | 10 | .2000 |
| | 20 | .3000 |
| | 25 | .3750 |
| | 30 | .6500 |
| | 40 | 1.3800 |
| | 50 | 2.1500 |
| | 60 | 2.8000 |
| | 70 | 4.5000 |
| | 75 | 5.1250 |
| | 80 | 5.5000 |
| | 90 | 6.0000 |

Como se observa, hay un 75% de sujetos que gastan S/.5.1250 como máximo en las compras del supermercado, y el 10% gasta S/.0.20 como máximo en la compra de artículos de aseo personal.

### I. Medidas de variación (dispersión)

Estudia la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados o dispersos. Existen diversas medidas de dispersión, entre las más utilizadas cabe destacar las siguientes:

- **Rango**. Mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el valor más bajo.
- **Varianza**. Mide la distancia existente entre los valores de la serie y la media. Se calcula como el sumatorio de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El sumatorio obtenido se divide por el tamaño de la muestra, y la varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

$$S_x^2 = \frac{\sum_{i=1}^{K}(Xi - \overline{X}_a)^2 fi}{n} = \frac{\sum_{i=1}^{n}(Xi - \overline{X}_a)^2}{n}$$

También:

$$S_x^2 = \frac{\sum_{i=1}^{K}(Xi - \overline{X}_a)^2 fi}{n-1} = \frac{\sum_{i=1}^{n}(Xi - \overline{X}_a)^2}{n-1}$$

- **Desviación típica**. Se calcula como la raíz cuadrada de la varianza.
- **Coeficiente de variación de Pearson**. Se calcula como cociente entre la desviación típica y la media.

## Ejemplo:

Utilice la serie de datos de la estatura de los alumnos de una clase y calcule sus medidas de dispersión.

| Variable | Frecuencias absolutas | | Frecuencias relativas | |
|---|---|---|---|---|
| (Valor) | Simple | Acumulada | Simple | Acumulada |
| 1,20 | 1 | 1 | 3,3% | 3,3% |
| 1,21 | 4 | 5 | 13,3% | 16,6% |
| 1,22 | 4 | 9 | 13,3% | 30,0% |
| 1,23 | 2 | 11 | 6,6% | 36,6% |
| 1,24 | 1 | 12 | 3,3% | 40,0% |
| 1,25 | 2 | 14 | 6,6% | 46,6% |
| 1,26 | 3 | 17 | 10,0% | 56,6% |
| 1,27 | 3 | 20 | 10,0% | 66,6% |
| 1,28 | 4 | 24 | 13,3% | 80,0% |
| 1,29 | 3 | 27 | 10,0% | 90,0% |
| 1,30 | 3 | 30 | 10,0% | 100,0% |

## Solución

- **Rango**. Diferencia entre el mayor valor de la muestra (1,30) y el menor valor (1,20). Luego el rango de esta muestra es 10 cm.

- **Varianza**. Recuerde que la media de esta muestra es 1,253; luego, aplicando la fórmula, se obtiene como varianza 0,0010.
- **Desviación típica**. Es la raíz cuadrada de la varianza. Entonces: 0.0320.
- **Coeficiente de variación de Pearson**. Se calcula como cociente entre la desviación típica y la media de la muestra. Por tanto:

$$Cv = 0{,}0320 / 1{,}253$$

$$Cv = 0{,}0255$$

Resultados con SPSS

**Estadísticos**

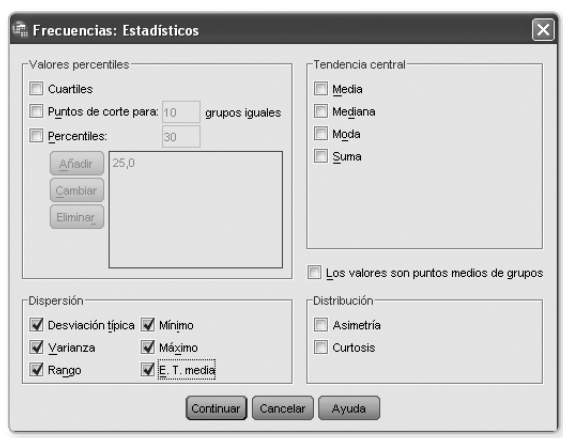
Estatura

| | | |
|---|---|---|
| N | Válidos | 30 |
| | Perdidos | 0 |
| Error típ. de la media | | ,00592 |
| Desv. típ. | | ,03241 |
| Varianza | | ,001 |
| Rango | | ,10 |
| Mínimo | | 1,20 |
| Máximo | | 1,30 |

### J. Medidas de variación con SPSS

Las medidas o estadísticas de variación se obtienen de forma muy similar, desde los mismos menús y los cuadros de diálogo que las de posición y tendencia central. Las más empleadas son la varianza y desviación típica, que aparecen por defecto en muchos menús de SPSS y en aquellas opciones que se denominan Descriptivas, sin especificar exactamente a qué se refieren.

La varianza y desviación típica, rango, mínimo y máximo pueden obtenerse a partir del procedimiento Frecuencia, en el botón Estadísticas. Desde aquí puede pedirse que se calcule la desviación típica, varianza y amplitud total o rango de las puntuaciones.




Otra forma de proceder para obtener medidas de variación es desde el menú Analizar > Estadísticos descriptivos > Descriptivos.

**Ejemplo:**

Para la venta de aseo personal, desde cualquiera de los cuadros de diálogo señalados, se obtienen los siguientes resultados:

**Statistics**

Venta de Artículos de Aseo Personal

| | | |
|---|---|---|
| N | Valid | 30 |
| | Missing | 0 |
| Std. Deviation | | 2.2483 |
| Variance | | 5.0550 |
| Range | | 6.30 |
| Minimum | | .20 |
| Maximum | | 6.50 |

Conviene tener en cuenta un aspecto importante. El programa SPSS tiene la peculiaridad de que proporciona la cuasivarianza; esto es, cuando se le pide la varianza y desviación típica, lo que da exactamente es el resultado de las siguientes fórmulas:

$$S'^2_x = \frac{\sum (X_i - \overline{X})^2}{n-1} \qquad S'_x = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

### K. Medidas de forma

Las medidas de forma permiten conocer qué forma tiene la curva que representa la serie de datos de la muestra. En concreto, se pueden estudiar las siguientes características de la curva:

#### a. Asimetría

El concepto de asimetría se refiere a la curva que forman los valores de la serie de datos; presenta la misma forma a la izquierda y derecha de un valor central (media aritmética), y se calcula de la siguiente manera:

$$A_S = \frac{\overline{X} - Mo}{S} = \frac{3(\overline{X} - Me)}{S}$$

Donde:

$\overline{X}$: Media aritmética

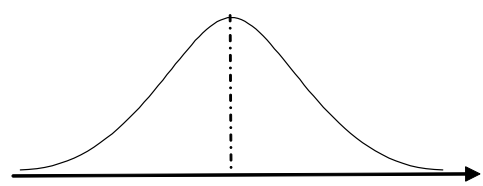
Mo: Moda

Me: Mediana

S: Desviación estándar o desviación típica

- **Simétrica**

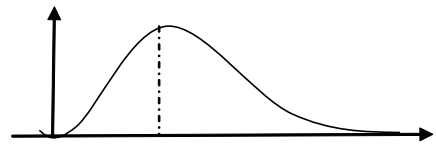
  Si $A_s = 0$:



$$\overline{X} = Me = Mo$$

Existe la misma concentración de valores a la derecha y a la izquierda de la media, tal como se observa en la figura. La media, la mediana y la moda tienen el mismo valor.

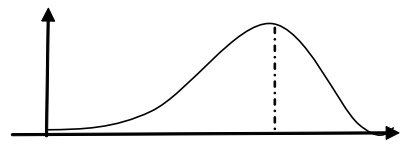
- **Simétrica a la derecha o positiva**
  Si $A_s > 0$.



$\overline{X} > Me > Mo$

Existe mayor concentración de valores a la derecha de la media que a su izquierda. La media es mayor que la mediana, y la mediana mayor que la moda.

- **Simétrica a la derecha o positiva**
  Si $A_s < 0$.



$\overline{X} < Me < Mo$

Existe mayor concentración de valores a la izquierda de la media que a su derecha. La media es menor que la mediana, y la mediana es menor que la moda.

### b. Curtosis

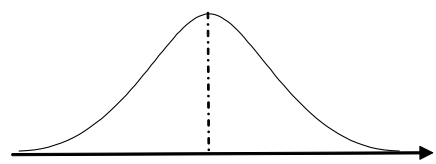
El coeficiente de curtosis analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución.

$$C_u = \frac{RI}{2\left[P_{90} - P_{10}\right]}$$

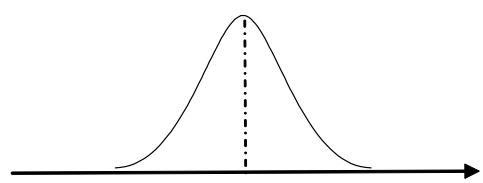
Donde:

RI es el rango intercuartílico y se calcula $RI = Q_3 - Q_1$. Se definen tres tipos de distribuciones, según su grado de curtosis:

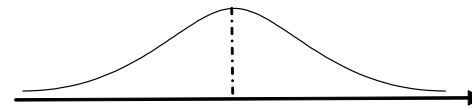
- **Distribución mesocúrtica**
  Si Cu=0.263, presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).



- **Distribución leptocúrtica**
  Si Cu < 0.263, presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

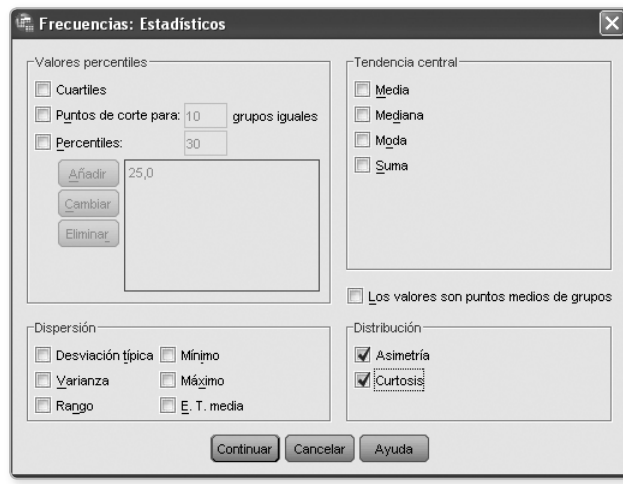


- **Distribución platicúrtica**
  Si Cu > 0.263, presenta un reducido grado de concentración alrededor de los valores centrales de la variable.



## L. Medidas de forma con SPSS

La asimetría (Skewness) y la curtosis (Kurtosis) pueden obtenerse a partir del menú Analizar > Frecuencias, pulsando el botón Estadísticas y luego marcando las respectivas distribuciones.




Se obtiene el mismo resultado desde el menú Analizar > Estadísticas descriptivos > Descriptivas. Desde ambos cuadros de diálogo se ofrecen exactamente las mismas opciones: el índice de asimetría de Fisher para cada variable y el índice de curtosis. Además del valor de la estadística, tanto para la asimetría como para la curtosis, el SPSS proporciona información sobre el error típico en el cálculo de estos índices. Vea un ejemplo para algunas variables de la encuesta en el supermercado:

**Statistics**

| | | Venta de Artículos de Aseo Personal | Venta de Verduras | Venta de Abarrotes |
|---|---|---|---|---|
| N | Valid | 30 | 30 | 30 |
| | Missing | 0 | 0 | 0 |
| Skewness | | .435 | .562 | -.050 |
| Std. Error of Skewness | | .427 | .427 | .427 |
| Kurtosis | | -1.413 | -.565 | .082 |
| Std. Error of Kurtosis | | .833 | .833 | .833 |

### M. Explorar

Las gráficas más importantes de este menú son las siguientes:

#### a. Diagrama de tallo y hojas

Otra de las representaciones que sirven para reflejar distribuciones de frecuencias es el *diagrama de tallo y hojas*. Para confeccionarlo, se utiliza el menú **Analizar > Estadísticos descriptivos > Explorar,** cuyo cuadro de diálogo es el siguiente:




Trasladar aquí la variable o variables para las que se desee obtener un análisis descriptivo.

Pulse aquí para solicitar un diagrama de tallo y hojas.

El cuadro de diálogo ofrece diferentes posibilidades de análisis de datos; sin embargo, solo se mencionarán algunas. En este caso, elabore un diagrama de tallo y hojas seleccionando las variables para las que se desee este tipo de diagrama, y se trasladan al cuadro Dependientes. Luego, seleccione el botón de la opción Gráficos y se mostrará el siguiente cuadro de diálogo:




Para diagramas de tallo y hojas, active esta casilla.

**Ejemplo:**

Con la variable Venta de artículos de aseo personal, se obtienen los siguientes resultados:

Venta de artículos de aseo personal Stem-and-Leaf Plot.

```
Frequency    Stem &  Leaf
    9.00      0 .  222233345
    5.00      1 .  02355
    4.00      2 .  0335
    2.00      3 .  03
    2.00      4 .  55
    4.00      5 .  0556
    4.00      6 .  0005
Stem width:     1.00
Each leaf:      1 case(s)
```

### b. Gráficos de cajas y bigotes

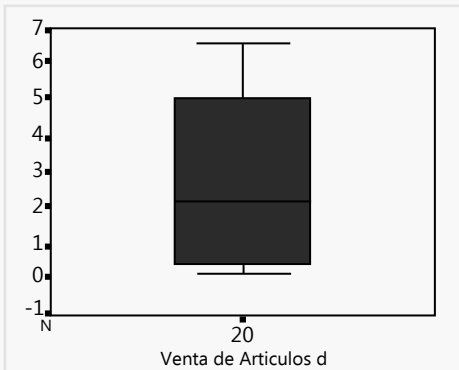
SPSS también ofrece la posibilidad de elaborar los gráficos que expresan la dispersión de los datos en una variable. En concreto, desde el menú **Analizar** > **Explorar** (Gráficos) puede definirse un diagrama de cajas y bigotes para una variable en estudio.



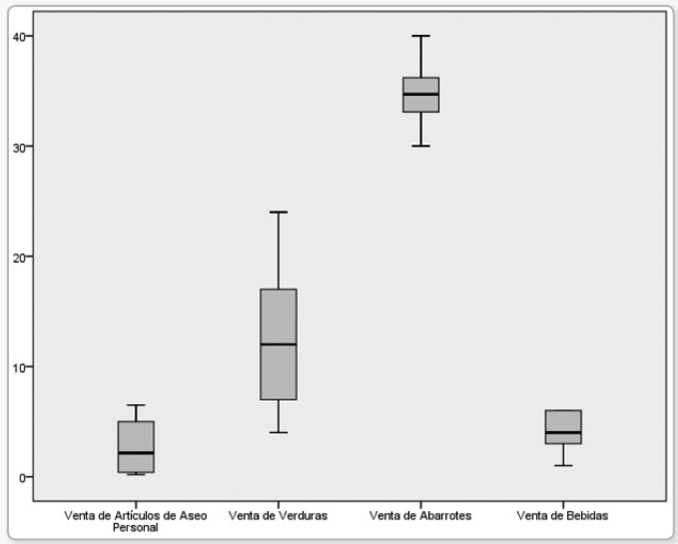

Por defecto, aparece selecionado Niveles de los factores juntos. Este tipo de diagramas ofrece información sobre el rango de la variable y los cuartiles.

**Ejemplo:**

Para la variable Artículos de aseo personal, se obtiene el gráfico siguiente:




Se observa en la gráfica que los gastos oscilan entre 0,20 y 6,50 nuevos soles, estando la mediana en 2,15 nuevos soles. Pero si selecciona Dependientes juntas, este tipo de diagramas ofrecerá información sobre el rango y los cuartiles de las variables selecionadas para el análisis; por ejemplo, para las variables Artículos aseo personal, Verduras, Abarrotes y Bebidas, se obtiene el gráfico siguiente:




### N. Tablas de contingencia

Observe cómo se obtienen distribuciones conjuntas de frecuencias y su representación gráfica para dos o más variables de diversos tipos. Los datos utilizados son de la encuesta del supermercado. La forma de proceder en el SPSS es tal como se ha observado anteriormente. Todos los análisis se realizan desde el menú Analizar > Estadísticas descriptivas > Tablas de contingencia (Crosstabs...).

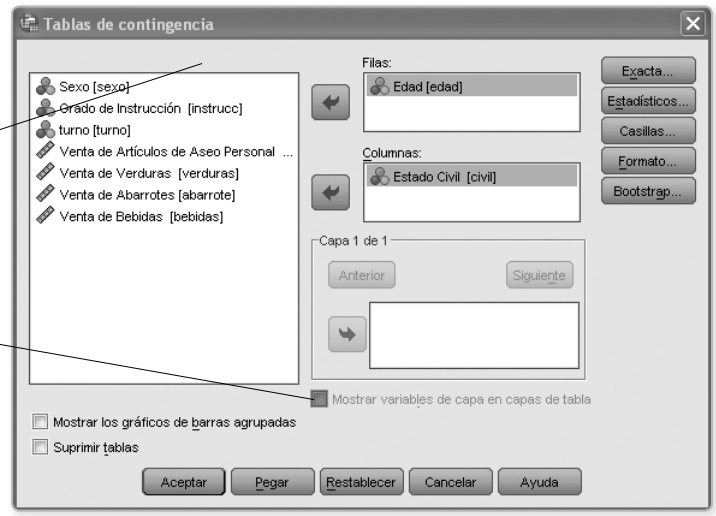



#### a. Dos variables cualitativas

Para describir las dos variables cualitativas, elabore una *tabla de contingencia (Crosstabs...)*. Si se selecciona el procedimiento **Tablas de contingencia**, el cuadro de diálogo que se muestra en la figura siguiente aparece.

Pulse para indicar si desea obtener los datos en porcentajes.

Seleccione si se desea obtener un diagrama de barras para las variables.



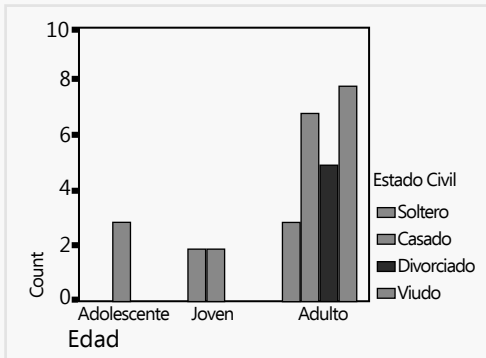

**Ejemplo:**

Conocer la distribución conjunta de frecuencias de las variables Edad y Estado civil. Para indicar al SPSS esta instrucción, seleccione las variables. Las categorías de la variable Edad pueden ser las filas (Row[s]) de la tabla de frecuencias, y las de la variable Estado civil, las columnas (Colum[s]) como se muestra en la figura anterior. Este cuadro de diálogo también permite obtener una representación gráfica para las dos variables. Si pulsa el botón Aceptar, el resultado obtenido es el siguiente:

**Edad * Estado Civil Crosstabulation**

Count

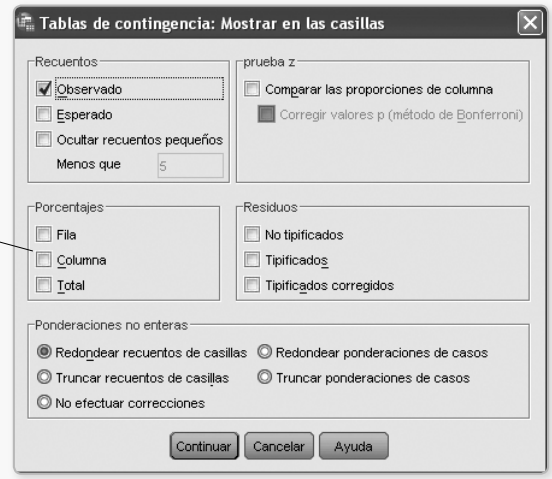
| | | Estado Civil | | | | Total |
|---|---|---|---|---|---|---|
| | | Soltero | Casado | Divorciado | Viudo | |
| Edad | Adolescente | | 3 | | | 3 |
| | Joven | 2 | 2 | | 1 | 5 |
| | Adulto | 3 | 7 | 4 | 8 | 22 |
| Total | | 5 | 12 | 4 | 9 | 30 |




En la tabla de resultados se observan las *distribuciones conjuntas y marginales* de las variables y la representación gráfica del diagrama de barras conjunto. Como se observa, de los 30 sujetos de la muestra, tres son adolescentes, cinco son jóvenes y 22 son adultos. De los adultos, la mayoría son viudos (8), etc.,

Las distribuciones conjuntas también pueden obtenerse en frecuencias relativas o porcentajes para ambas variables (filas y columnas). Para ello, seleccione el botón Celdas (Cells) del cuadro de diálogo anterior, cuyo aspecto es el siguiente:

Señalar si se desea obtener las frecuencias conjuntas relativas para las filas y las columnas.




Seleccionando las opciones marcadas, se obtienen los mismos resultados que antes, pero expresados en porcentajes.

**Edad * Estado Civil Crosstabulation**

| | | | Estado Civil | | | | |
|---|---|---|---|---|---|---|---|
| | | | Soltero | Casado | Divorciado | Viudo | Total |
| Edad | Adolescente | Count | | 3 | | | 3 |
| | | % within Edad | | 100.0% | | | 100.0% |
| | | % within Estado Civil | | 25.0% | | | 10.0% |
| | Joven | Count | 2 | 2 | | 1 | 5 |
| | | % within Edad | 40.0% | 40.0% | | 20.0% | 100.0% |
| | | % within Estado Civil | 40.0% | 16.7% | | 11.1% | 16.7% |
| | Adulto | Count | 3 | 7 | 4 | 8 | 22 |
| | | % within Edad | 13.6% | 31.8% | 18.2% | 36.4% | 100.0% |
| | | % within Estado Civil | 60.0% | 58.3% | 100.0% | 88.9% | 73.3% |
| Total | | Count | 5 | 12 | 4 | 9 | 30 |
| | | % within Edad | 16.7% | 40.0% | 13.3% | 30.0% | 100.0% |
| | | % within Estado Civil | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% |

En el caso de los solteros, del 100% de ellos se observa que el 0% son adolescentes, el 40% son jóvenes y el 60% son adultos (la lectura fue en columna). En el caso de los adultos, del 100% se observa que el 13.6% son solteros, el 31.8% son casados, el 18.2% son divorciados y el 36.4% son viudos (la lectura fue en Fila).

Las tablas de contingencia solo tienen sentido para variables de tipo cualitativo (nominal u ordinal). Si desea representar la distribución conjunta de dos variables cuantitativas, agrupe los valores de cada una de las variables en intervalos de la misma amplitud (para ello, puede usarse el menú Transformar > Recodificar). Una vez obtenidos los intervalos para dichas variables, puede elaborarse la tabla de contingencia siguiendo el mismo procedimiento que si las variables fueran cualitativas.

La gráfica más adecuada para dos variables cuantitativas es el diagrama de dispersión:

- **Una variable cualitativa y otra cuantitativa**. Si se quiere cruzar una variable cualitativa con una cuantitativa, se procede igual que para variables cualitativas, excepto que para resumir los datos de la cuantitativa hay que emplear estadísticos (sumas, medias, etc.). También se puede representar gráficamente la relación entre estas dos variables. La presentación gráfica es más recomendable porque se puede analizar con mayor claridad las variables en estudio.
- **Una variable cuantitativa y dos cualitativas**. Es frecuente la representación de una variable dependiente cuantitativa, en función de dos independientes de tipo cualitativo. En estos casos, también se puede representar gráficamente la relación entre estas tres variables. En forma similar a la anterior, la presentación gráfica es más recomendable porque se pueden analizar con mayor claridad las variables en estudio (ver § 9).

### 4.2.2. EJERCICIO PRÁCTICO 2

Los datos aparecen recopilados en las variables ***Día1***, ***Día2*** y ***Día3,*** que indican el número de horas trabajadas en los tres días; además, se ha recopilado la variable ***Sexo,*** que toma valores 0 para mujeres y 1 para hombres; y la variable ***Edad***: 1 para adolescentes, 2 para jóvenes y 3 para adultos. Finalmente, se incluyen tres datos relacionados con las terapias que ha recibido el sujeto. Todas ellas toman 1, si el sujeto ha recibido la terapia y 0, en caso contrario. Las variables de terapia son ***Estrés***, ***Ansiedad*** y ***Fobia***.

Los datos representan a 10 sujetos que participaron en el estudio, cuyos resultados obtenidos se muestran en la tabla siguiente.

| Sujeto | Día1 | Día2 | Día3 | Sexo | Edad | Estrés | Ansiedad | Fobia |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.7 | 6.3 | 6.8 | 0 | 1 | 0 | 1 | 0 |
| 2 | 4.8 | 5.4 | 4.3 | 1 | 3 | 0 | 0 | 0 |
| 3 | 7.6 | 7.9 | 8.6 | 0 | 2 | 0 | 0 | 0 |
| 4 | 5.7 | 6.0 | 6.7 | 1 | 1 | 0 | 0 | 1 |
| 5 | 3.8 | 4.6 | 4.9 | 1 | 3 | 0 | 0 | 0 |
| 6 | 7.5 | 8.0 | 8.5 | 1 | 2 | 0 | 1 | 0 |
| 7 | 6.4 | 7.0 | 7.3 | 1 | 1 | 1 | 1 | 0 |
| 8 | 7.7 | 8.1 | 8.5 | 0 | 2 | 0 | 0 | 0 |
| 9 | 4.0 | 4.5 | 5.0 | 0 | 3 | 1 | 0 | 0 |
| 10 | 5.7 | 6.1 | 6.8 | 1 | 2 | 1 | 0 | 1 |

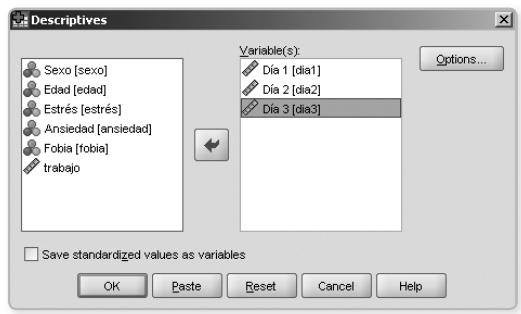
#### A. Estadísticas descriptivas

Cuando se realizan análisis descriptivos de los datos se utiliza el procedimiento Descriptivas, que proporciona distintas estadísticas de tendencia central, dispersión, asimetría y curtosis. Además, permite obtener puntuaciones típicas y guardarlas como nuevas variables.

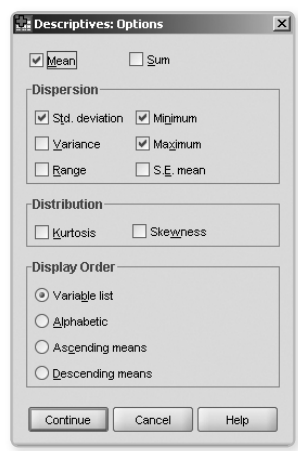
Para seleccionar el procedimiento descriptivo, se deben escoger las opciones del menú Analizar > Estadísticas > Descriptivos:




A continuación, se muestra un cuadro de diálogo donde se listan las variables del archivo de datos. Para indicar a SPSS cuáles son las variables que se desea analizar, se seleccionan dichas variables de la lista pulsando sobre su nombre con el mouse. Luego, se pulsa sobre el botón Flecha para trasladar las variables al cuadro de variables seleccionadas.




Siempre los cuadros de diálogo de procedimientos de SPSS poseen varios botones. Los botones comunes son Aceptar (OK), Pegar, Restablecer, Cancelar y Ayuda. El cuadro de diálogo del procedimiento Descriptivos contiene el botón específico denominado Opciones... Los tres puntos significan que, pulsando sobre este botón, aparecerá un subcuadro de diálogo que permite seleccionar los estadísticos que se desean calcular.



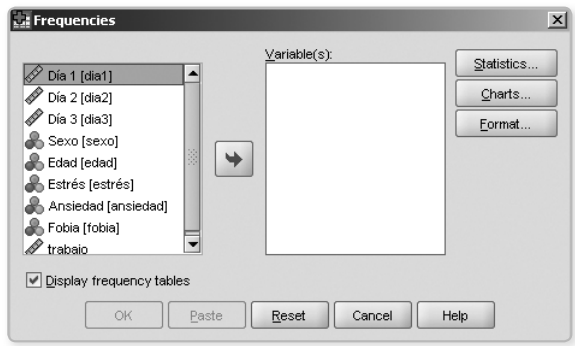
Para poder realizar los cálculos estadísticos, pulse el botón OK del cuadro de diálogo del procedimiento Descriptivos. SPSS muestra los cálculos solicitados en la ventana Visor de resultados.

**Descriptive Statistics**

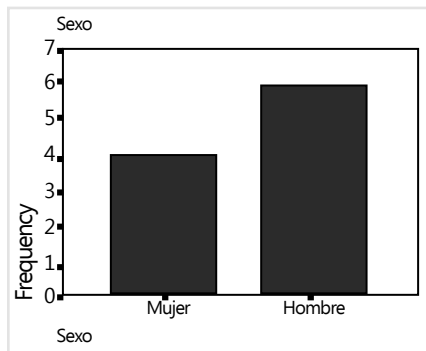
| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Día 1 | 10 | 3.80 | 7.70 | 5.8900 | 1.4255 |
| Día 2 | 10 | 4.50 | 8.10 | 6.3900 | 1.3404 |
| Día 3 | 10 | 4.30 | 8.60 | 6.7400 | 1.5756 |
| Valid N (listwise) | 10 | | | | |

**B. Frecuencia**

Este procedimiento proporciona la distribución de frecuencias de una variable en forma de tabla y en forma de gráfica, además de varias estadísticas descriptivas. Para ejecutar este procedimiento, seleccione el menú **Analizar** > **Estadísticas descriptivas** > **Frecuencias**. Luego se muestra un cuadro de diálogo en el cual se indican las variables del archivo de datos que se está trabajando. Al igual que en el procedimiento Descriptivos, primero traslade al recuadro de la derecha (Variables) aquellas variables que desea analizar.

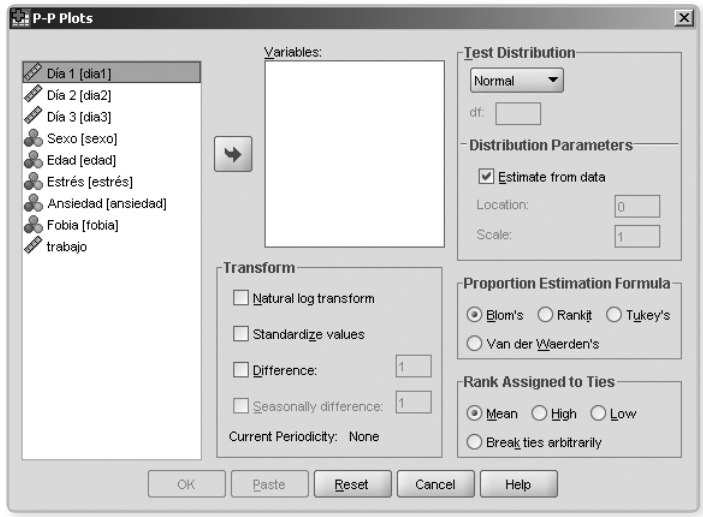



Si pulsa sobre el botón Estadísticas, pueden seleccionarse diversas estadísticas, que pueden ser de tendencia central y no central, dispersión, forma de distribución, percentiles, deciles, etc. En este cuadro de diálogo, también se puede utilizar el botón Gráficos (Charts...) para indicar a SPSS que se desea realizar un gráfico de barras, de sectores o un histograma; por ejemplo, si se quiere utilizar la gráfica de barras en la variable Sexo. El resultado se observa en el diagrama siguiente:



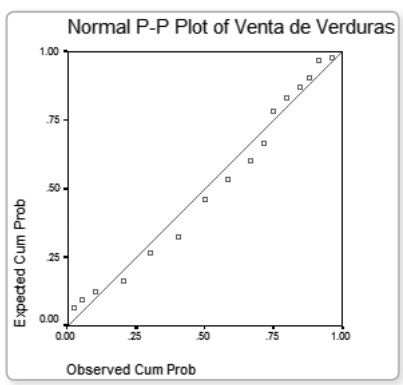
### C. P-P Plots

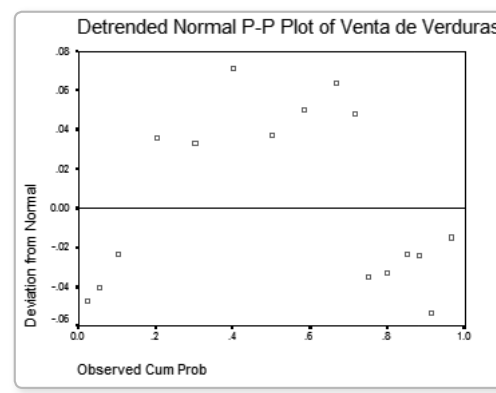
Para crear un gráfico de P-P, el cuadro de diálogo del gráfico deseado debe ser el siguiente:




 crea un gráfico de las proporciones acumuladas de una variable, respecto a las de una distribución cualquiera de prueba; es decir, por cada variable especificada, proporciona dos gráficos: en el primero, se representan los valores de la función de distribución acumulativa esperada, bajo el supuesto de normalidad (por defecto) frente a los observados; en el segundo, se representan los residuos. Los gráficos de probabilidad suelen emplearse para determinar si la distribución de una variable coincide con una distribución dada.

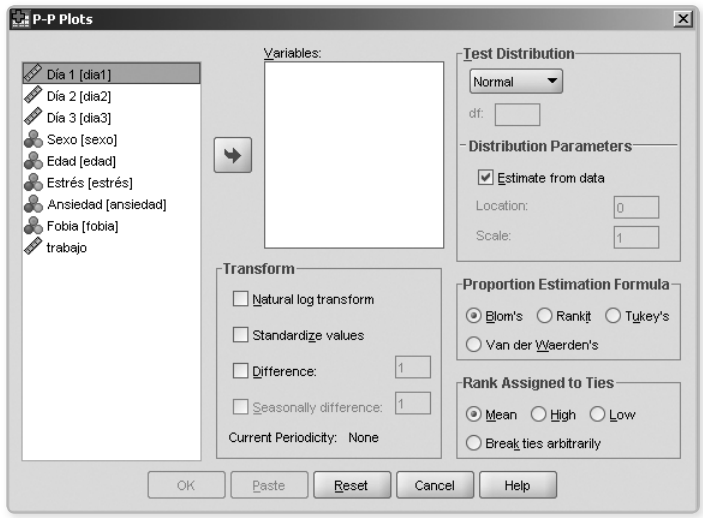
Si la variable seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta. Entre las distribuciones de prueba disponibles se encuentran beta, Chi-cuadrado, exponencial, gamma, seminormal, Laplace, logístico, Lognormal, normal, Pareto, t de Student, Weibull y uniforme. Según la distribución elegida, pueden especificarse distintos grados de libertad y otros parámetros. De modo que, para la variable Venta de Verduras, se puede decir que sí tiene una distribución normal.



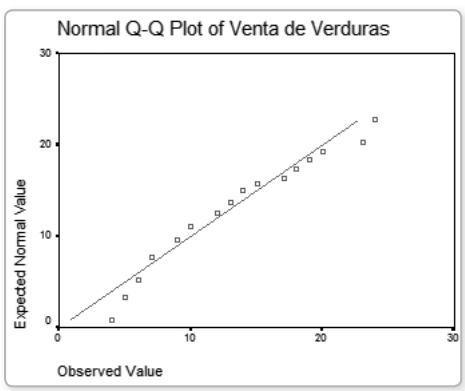
**D. Q-Q Plots**

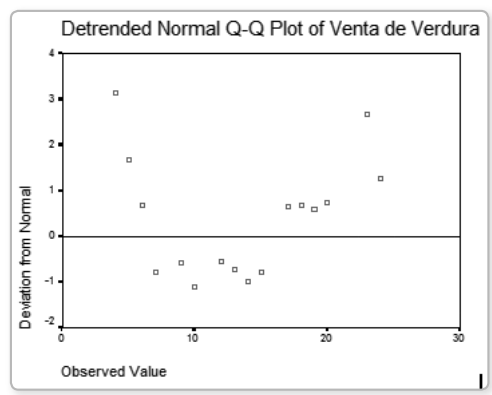
Para crear un gráfico de Q-Q, el cuadro de diálogo del gráfico deseado debe ser como sigue:




 crea un gráfico con los cuantiles de distribución de una variable, con relación a los cuantiles de una distribución cualquiera de prueba. En otras palabras, por cada variable especificada, se proporciona dos gráficos: en el primero se representan los valores de los cuantiles esperados bajo el supuesto de distribución normal (por defecto) de media cero y varianza uno frente a los observados; en el segundo, en lugar de los cuantiles esperados, se representan los residuos. Las distribuciones de prueba disponibles son las mismas que en P-P.

Así, en el ejemplo para la variable venta de verduras, se puede decir que sí tiene una distribución normal.



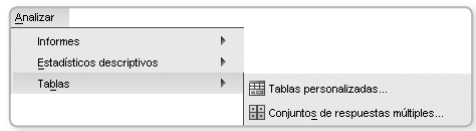
**EJERCICIOS PROPUESTOS**

1. Trabajando los datos de la encuesta del supermercado, resuelva las siguientes preguntas:
   a) ¿Cuál es el gasto mínimo y máximo para cada una de las variables (aseo abarrotes, etc.) de los sujetos de la muestra?
   b) ¿Qué porcentaje gasta menos de S/.3.00 en cada una de las variables?
   c) ¿Cuál es el valor central de cada variable?
   d) Obtenga el valor de la venta de artículos de aseo, que es superado por el 30% de los sujetos de la muestra.
   e) ¿Cuántos sujetos tienen grado de instrucción primaria?

2. Trabajando con los datos del ejercicio 2, terapias a los trabajadores, responda las siguientes preguntas:
   a) ¿Cuántas horas trabajan como mínimo el 70% de los sujetos?
   b) Obtenga la representación gráfica más adecuada para las siguientes variables: sexo, edad, horas trabajadas.
   c) ¿Qué porcentaje de sujetos trabaja más de 3 horas al día?
   d) Confeccione el diagrama de tallo y hojas en la variable Horas trabajadas, para la muestra de varones, para la de mujeres y para la muestra total. Comente el resultado.
   e) ¿Quiénes trabajan más horas: los varones o las mujeres?

## 4.3. TABLAS PERSONALIZADAS

En primer lugar, se establecerán criterios generales sobre su estructura, que se deben tener en cuenta en la construcción de cuadros. Del mismo modo, se estudiarán los tipos de cuadros que se pueden elaborar para distintos fines y de acuerdo con las características de los datos disponibles.




### A. Tablas personalizadas...

Genera tablas que muestran estadísticos de clasificación cruzada y de subgrupo. Las variables de resumen seleccionadas se resumen en las casillas definidas por las variables de subgrupo; para todas las variables resumidas, se emplearán los mismos estadísticos. Si no existen variables de resumen, se muestran los recuentos. Se pueden definir casillas para mostrar subgrupos de casos empleando las combinaciones: Hacia abajo de la página (como filas distintas), A través de la página (como columnas distintas). La combinación de Hacia abajo y A través genera una presentación en forma de tabla de contingencia, repartidas en Tablas distintas. Las variables de esta lista subdividen la tabla en capas o en grupos de capas, de manera que solo puede verse una capa de la tabla cada vez. Una vez en el Visor, se debe pivotear para ver las otras capas. **Anidar** significa que se muestran las categorías de una variable bajo cada una de las categorías de la variable anterior. La apelación muestra las categorías de cada variable como un bloque, que se puede entender como la extracción de tablas diferentes y su unión en la misma presentación.

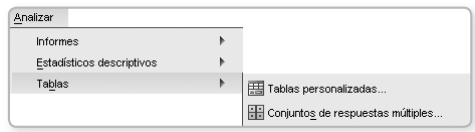
Se pueden generar tablas que muestren diferentes estadísticos para distintas variables (variables de respuestas múltiples, anidación y apilación mixta o totales complejos). En la mayoría de las circunstancias, es necesario seleccionar una variable de fila; al mismo tiempo, se puede seleccionar una o más variables para definir columnas o capas. Por ejemplo, para generar una tabla de contingencia simple, se debe seleccionar una variable para definir las filas y otra para definir las columnas. Si tiene una variable de capa, también es necesario tener una variable de fila y una de columna.

### B. Tablas de respuestas múltiples

Crea tablas de frecuencia y de contingencia básicas en las que una o más variables son un conjunto de respuestas múltiples (si no, será conveniente utilizar Tablas básicas). Si no se selecciona la opción Anidar para una dimensión, se producirán tablas separadas para cada variable, pero conectadas físicamente (concatenadas). Las restantes opciones son similares a las descritas anteriormente. Genera tablas especiales que contienen varias variables con los mismos valores; por defecto, las variables forman columnas y las categorías, filas (todas las variables de Frecuencias para deben tener las mismas categorías). Cada casilla muestra el número de casos en esa categoría; asimismo, permite variables de subgrupo mediante Subgrupos, y se puede dividir la tabla en columnas dentro de cada tabla o en tablas diferentes (capas). Al dividirla, se mostrarán todas las variables para cada subgrupo. Hay diversas variables en una de las listas de la sección Subgrupos; se puede elegir entre anidarlas o apilarlas (se crean las tablas por separado; pero se unen en la presentación).

### 4.3.1. EJERCICIO PRÁCTICO 1

Permitirá utilizar los principales procedimientos de tablas. Los datos empleados serán los recopilados en la encuesta del supermercado (ver § 2). Empiece a usar este menú haciendo clic en Analizar > Tablas (Tables), como se muestra en la siguiente figura:

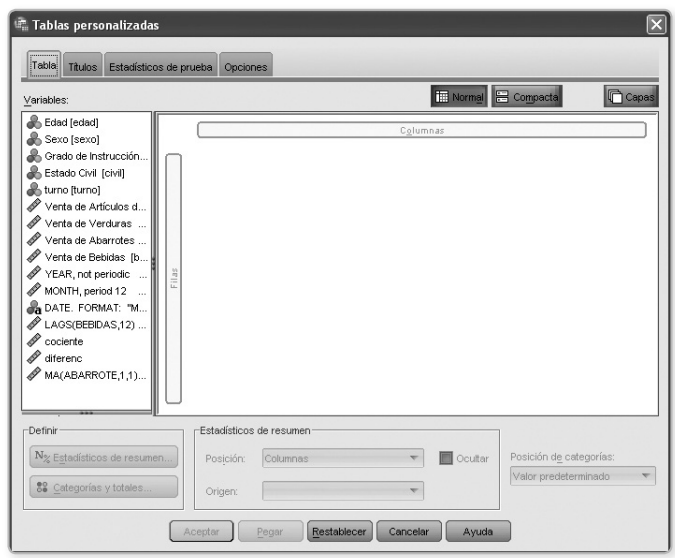



Para realizar un análisis con los procedimientos de Tablas (Tablas personalizadas, Tablas de respuestas múltiples), se deben seguir los cuadros de diálogo que se van presentado.

#### A. Tablas personalizadas

Para entenderlo mejor, divida Custom table en Normal, Compct y tablas con Layers (Capas).

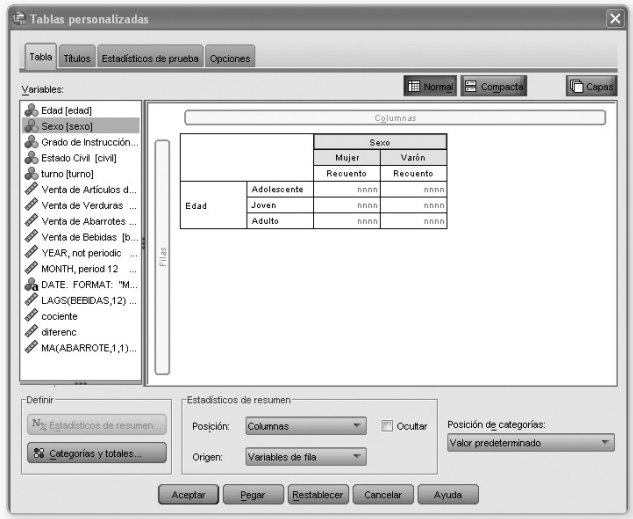
- **Normales (Normal).** Cuando se hace clic en Tablas generales (custom table), se muestra el cuadro de diálogo que ofrece el siguiente aspecto:




En el cuadro izquierdo aparece el listado de variables del archivo de datos que está abierto. El cuadro de la derecha presenta dos botones: uno para las filas y el otro para las columnas, esperando que le indique qué variables se desea analizar. Las variables se trasladan de un cuadro a otro arrastrando.

**Ejemplo:**

Trasladando las variables al cuadro de la derecha del modo siguiente: la variable Edad a la fila y la variable Sexo a la columna, obteniéndose la siguiente presentación:

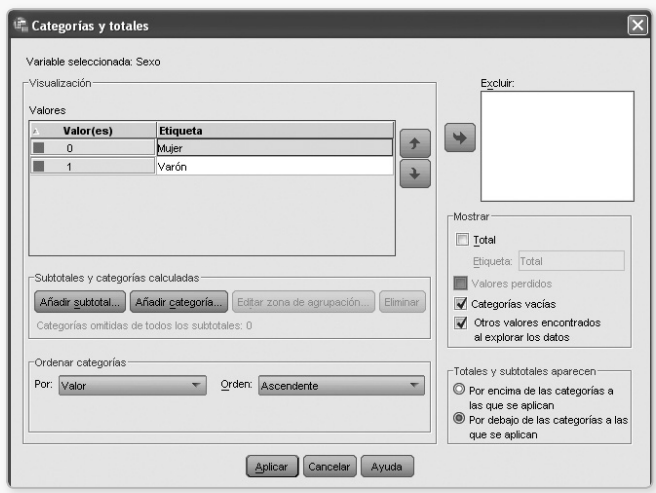



Luego pulse el botón Aceptar (OK) para que se ejecute la orden. Los resultados obtenidos aparecen en la ventana Visor de resultados:

**Table 1**

| | | Sexo | |
|---|---|---|---|
| | | Mujer | Varón |
| | | Count | Count |
| Edad | Adolescente | 3 | 0 |
| | Joven | 4 | 1 |
| | Adulto | 19 | 3 |

Además, si pulsa el botón **Categories and totals**, se obtiene como resultado el cuadro de diálogo siguiente:



Aquí se pueden excluir las categorías (cuadro izquierdo) que usted desee, solo debe trasladar las categorías al cuadro del lado derecho con el botón  .

Si la categoría Joven se traslada al cuadro de la derecha, excluyéndola del análisis, el resultado será el siguiente:

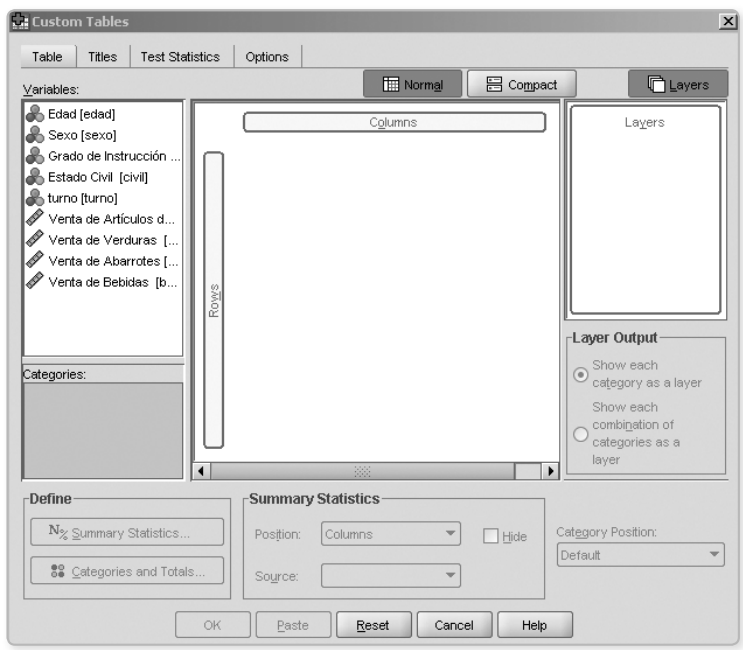
**Table 1**

| | | Sexo | |
|---|---|---|---|
| | | Mujer | Varón |
| | | Count | Count |
| Edad | Adolescente | 3 | 0 |
| | Adulto | 19 | 3 |

Para el caso de **Compct**, es similar a **Normal**.

### B. Layers (Capas)

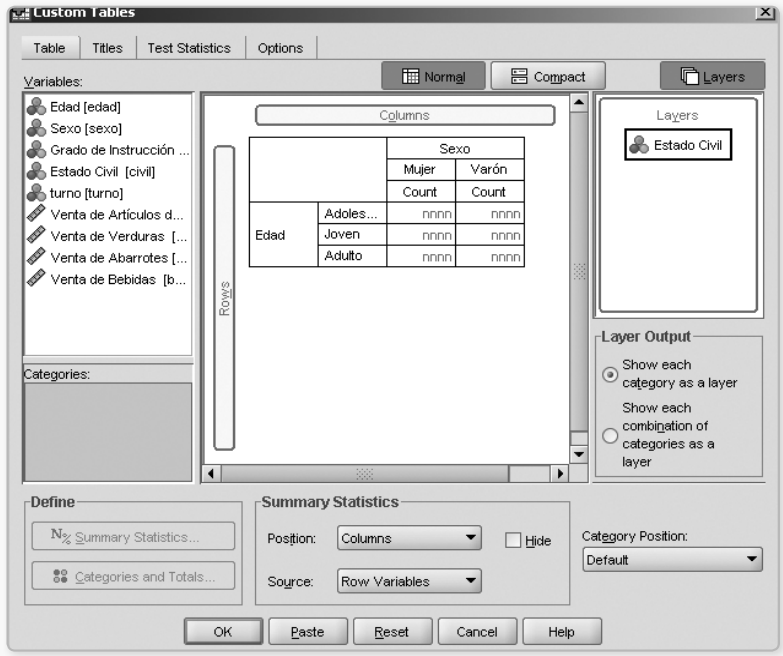
Si hace clic en Layers cuando se está en el cuadro de diálogo Tablas generales (custom table), el cuadro siguiente aparece:




En el cuadro izquierdo se muestra el listado de variables del archivo de datos que está abierto. El primer cuadro de la derecha presenta dos botones: uno para las filas y el otro para las columnas; por otro lado, el siguiente cuadro presenta el botón Layers. Estos cuadros están esperando la indicación de las variables a analizar, y las variables se trasladan arrastrando de un cuadro a otro.

**Ejemplo:**

Trasladando las variables al cuadro de la derecha del modo siguiente: la variable Edad a la fila, la variable Sexo a la columna y la variable Estado civil a Lager, se obtiene la siguiente presentación:




Por ejemplo, trasladando las variables al cuadro de la derecha del modo siguiente: la variable «Edad» a la fila, la variable «Sexo» a la columna y la variable «Estado civil» a «Lager». Se obtiene el siguiente resultado:

**Table 1**

Estado Civil Soltero

| | | Sexo | |
|---|---|---|---|
| | | Mujer | Varón |
| | | Count | Count |
| Edad | Adolescente | 0 | 0 |
| | Joven | 2 | 0 |
| | Adulto | 2 | 1 |

En la misma tabla, se pueden analizar los subgrupos de la variable Estado civil (soltero, casado, divorciado y viudo); solo haga doble clic en Estado civil: Soltero, como se muestra en la siguiente tabla:



### 4.3.2. EJERCICIO TEÓRICO

En el ejercicio práctico, se mostrarán los lineamientos para la construcción de cuadros estadísticos. La figura muestra un cuadro típico que reúne un conjunto de características, que son las siguientes:

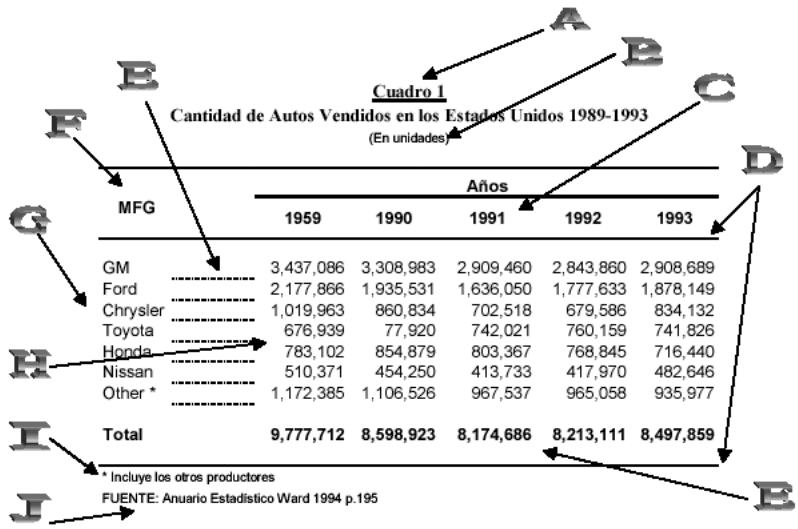



*Figura 4.1 Partes de un cuadro estadístico*

**A.** Es el número de cuadro y título (A). Todo cuadro estadístico debe ser numerado.

**B.** Las unidades (B) en que están medidas las cantidades presentadas.

**C.** Los nombres colocados en la parte superior de las columnas se llaman «cabeceras» (C).

**D.** También se ponen líneas continuas y de mayor intensidad para separar las cabeceras de columnas (D).

**E.** Para cuadros muy extensos, es conveniente poner líneas indicadoras (E) que permitan una lectura fácil.

**F.** Constituyen los niveles de la variable respuesta (F).

**G.** Los nombres colocados al lado izquierdo de las columnas se llamarán «marcas de categorías» (G).

**H.** Los datos se pueden ordenar descendente o ascendentemente (H).

**I.** Las notas al pie del cuadro proveen explicaciones adicionales que deberían estar incorporadas en el cuadro, y permitan esclarecer aún más la información (I).

**J.** La fuente (J) es una nota importante que forma parte del cuadro.

#### A. Tablas cruzadas

Siempre es importante saber la relación que existe entre dos o más variables. Por ejemplo, en el ámbito de la investigación de mercados, estará interesado en saber los siguiente:

- ¿Cuántos compradores de determinado producto son hombres?
- ¿El nuevo producto se ajusta a la edad y nivel educativo?
- ¿Está relacionada la propiedad del producto con los gastos realizados?

Estas preguntas se pueden resolver mediante el estudio de las tablas cruzadas, porque permite analizar diversas variables simultáneamente, mostrando la distribución conjunta de dos o más variables con un número limitado de categorías o distintos valores.

### B. Dimensiones de una tabla

Los cuadros pueden ser bidimensionales: una variable en las columnas y la otra variable en las filas. Una tercera variable definirá una distribución conjunta para cada uno de los valores que tome. Así, en esta tercera dimensión se definirán segmentos. Como muestra la figura 4.2, las tres dimensiones de una tabulación cruzada se asemejan a las tres dimensiones físicas: ancho, profundidad y altura.

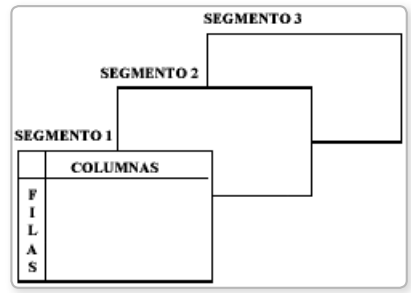



*Figura 4.2 Tres dimensiones para una tabulación cruzada*

### C. Tablas bidimensionales

Es la tabulación cruzada bivariada conocida como «tabla bidimensional». Dos variables han sido clasificadas de manera cruzada, y los porcentajes pueden calcularse tanto por filas como por columnas. La manera correcta de calcular los porcentajes depende de cuál es la variable independiente y cuál es la dependiente.

| Familiaridad | Tiempo de Resistencia | | |
|---|---|---|---|
| | Menos de 13 años | De 13 a 30 años | Mas de 30 años |
| No familiar | 45 | 34 | 55 |
| familiar | 52 | 53 | 27 |
| total | 97 | 87 | 82 |

*Figura 4.3 Tiempo de residencia y familiaridad con el departamento de almacén.*

En la figura 4.3, la variable independiente puede ser el tiempo de residencia y la familiaridad; en cambio, la variable dependiente se da porque se está interesado en conocer el grado de familiaridad que los residentes de una zona tienen con respecto a un nuevo almacén de ventas.

Al ser el Tiempo de residencia la variable independiente, la manera correcta de calcular los porcentajes será como se muestra en la figura.

| Familiaridad | Tiempo de Resistencia | | |
|---|---|---|---|
| | Menos de 13 años | De 13 a 30 años | Mas de 30 años |
| No familiar | 46.40% | 39.10% | 67.10% |
| familiar | 53.60% | 60.98% | 32.90% |
| total | 100.00% | 100.00% | 100.00% |

*Fig. 4. Familiaridad con el Departamento de Almacén por Tiempo de residencia*

Analizando el cuadro, se observa que las personas que viven en la zona entre 13 y 30 años, son los que están más familiarizados con el nuevo almacén. Sin embargo, se nota que la proporción de personas que viven en la zona por más de 30 años es menor que el grupo de gente que vive menos de 13 años.

### D. Tablas tridimensionales

Frecuentemente, la introducción de una tercera variable clarifica la asociación inicial observada entre dos variables. Al ingresar una tercera variable en el análisis, puede suceder cualquiera de las cuatro posibilidades:

- Clarifica la asociación observada entre las dos variables originales.
- Puede indicar falta de asociación entre las dos variables, aunque la asociación fue inicialmente observada.
- Puede indicar una asociación entre las dos variables iniciales, aunque la asociación inicialmente no fue observada.
- Puede indicar estabilidad en la asociación inicial.

| Compran Automóviles caros | Ingreso | | | |
|---|---|---|---|---|
| | Bajos | | Altos | |
| | Educación | | Educación | |
| | Superior | Secundarios | Superior | Secundarios |
| Si | 21% | 19% | 41% | 40% |
| No | 79% | 81% | 59% | 60% |
| Total | 100% | 100% | 100% | 100% |
| Encuestados | 100 | 700 | 150 | 50 |

*Figura 4.5 Propietarios de automóviles caros por grado de instrucción y nivel de ingresos*

CAPÍTULO

5

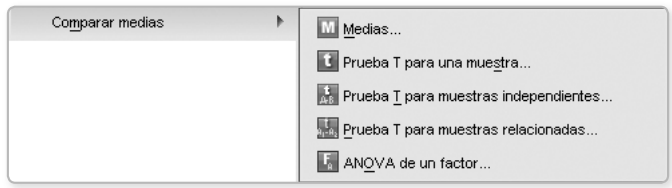
# COMPARAR MEDIAS, MODELO LINEAL GENERAL Y CORRELACIÓN

En este capítulo, eminentemente práctico, se presentan numerosas técnicas estadísticas básicas, como *comparar medias, modelo lineal general y correlación*; ademas, se discutirán situaciones en las que son aplicables cada una de ellas, las hipótesis previas necesarias para resolverlas y cómo realizar e interpretar los resultados.

Los temas que se presentan en este capítulo son: prueba T (para una muestra, muestras independientes, muestras relacionadas), ANOVA de un factor, ANOVA A-EF-CA, comparaciones múltiples entre medias, comparaciones a posteriori, análisis de varianza de un factor con medidas repetidas, y ANOVA A-EF-MR.

## 5.1. COMPARAR MEDIAS (PRUEBA DE HIPÓTESIS)

Comparar medias de SPSS es un procedimiento estadístico que permite realizar distintos tipos de *contrastes sobre una, dos y hasta "n" medias.*

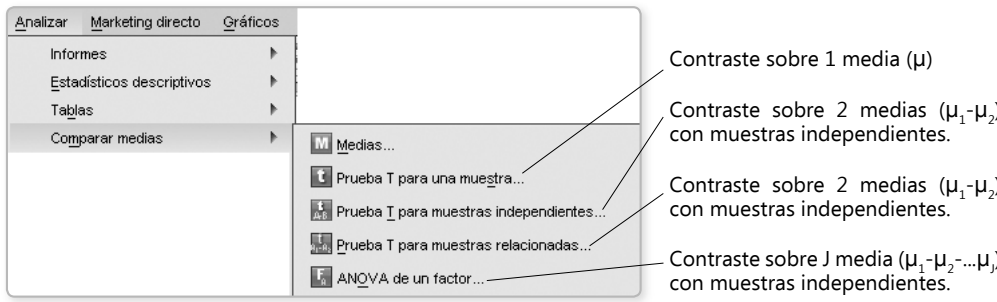



- **Medias**. Calcula medias de subgrupo y estadísticos univariados relacionados para variables dependientes, dentro de las categorías de una o más variables independientes. Permite mejorar en la creación de medias armónicas y geométricas. Pueden obtenerse el análisis de varianza de un factor, la eta y eta2, la prueba de linealidad con r y r2, y cálculo de estadísticas para la muestra total.
- **Prueba T para una muestra**. Prueba T de una muestra para comparar su media con la media de referencia de su elección; es decir, contrasta si la media de una sola variable difiere de una constante especificada.
- **Prueba T para muestras independientes**. Compara las medias de dos grupos de casos. Los sujetos deben asignarse aleatoriamente a dos grupos, de forma que cualquier diferencia en la respuesta sea debida al tratamiento (o falta de tratamiento) y no a otros factores. Cuando esto no es posible (por ejemplo, comparando una variable entre hombres y mujeres (ya que el sexo de una persona no se asigna aleatoriamente), asegúrese que las diferencias en otros factores no enmascaren o resalten una diferencia significativa entre las medias (por ejemplo, los estudios respecto al sexo). Además, compara estadísticas de muestras independientes, las medias de dos muestras para estimaciones combinada y separada de la varianza con la prueba de Levene para varianzas iguales.
- **Prueba T para muestras relacionadas**. Compara las medias de dos variables de un solo grupo, calcula las diferencias entre los valores de las dos variables de cada caso y contrasta si la media difiere de cero; además, compara estadísticas de muestras relacionadas: correlación entre pares, diferencia entre medias, probabilidad bilateral para prueba de no diferencia y para prueba de correlación cero entre pares.
- **ANOVA de un factor**. Genera un análisis de varianza de un factor para una variable dependiente cuantitativa, respecto a una única variable de factor (la variable independiente). El análisis de varianza se utiliza para contrastar la hipótesis de que varias medias son iguales. Esta técnica es una extensión de la prueba T para dos muestras. Además de determinar que existen diferencias entre las medias, es posible que se desee saber qué medias difieren. Existen dos tipos de contrastes para comparar medias: los contrastes a priori y las pruebas post hoc. Los contrastes a priori se plantean antes de ejecutar el experimento y las pruebas post hoc se realizan después de haber llevado a cabo el experimento. También se pueden contrastar las tendencias existentes a través de las categorías.
- **Contrastes**. Lineal, cuadrática, cúbico, de órdenes superior, definido por el usuario.
- **Pruebas de postc-hoc**. Duncan, LSD, Bonferroni, Student-Newman_Keuls, Scheffe, prueba alterna deTukey, Tukey-HSD, Sidak, Gabriel, hochberg, Tamhane, Dunnett T3, Waller-Duncan y otros.

### 5.1.1. EJERCICIO PRÁCTICO 1

Permitirá utilizar los principales procedimientos del menú Comparación de medias. Los datos empleados serán los recopilados en la encuesta del supermercado.

En el menú Analizar se encuentra el procedimiento Comparar medias, que permite realizar diferentes tipos de contrastes sobre una, dos y hasta "n" medias. El aspecto de dicho menú es el siguiente: Analizar > Comparación de medias (Compare Means).




Para realizar un análisis con los procedimientos de tablas personalizadas (medias, prueba T para una muestra, prueba T para muestras independientes, prueba T para muestras relacionadas, ANOVA de un factor), se deben seguir los cuadros de diálogo que se presenta.

A continuación, se brinda una explicación detallada del cuadro de diálogo correspondiente a los principales procedimientos.

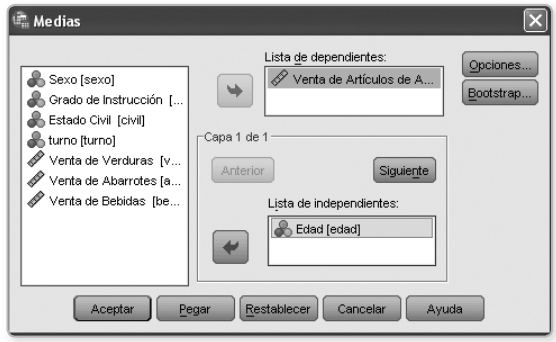
#### A. Medias

Utilizando el procedimiento Medias de SPSS, se pueden calcular medias y variancias de cada atributo de una determinada variable seleccionada. El cuadro de diálogo de la herramienta estadística **Medias** ofrece el siguiente aspecto:




En el cuadro de la izquierda se muestra el listado de las variables que tiene el archivo de datos que está abierto. Los cuadros de la derecha están vacíos, pues están esperando a que se le indique qué variable se quiere analizar. Dichas variables se trasladan de un cuadro a otro mediante el botón  .

En este ejemplo, se trabajará con las variables de la encuesta del supermercado. Para ello, traslade las variables que deba analizar (Venta de artículos de aseo y Edad) a los cuadros de la derecha del modo siguiente:




Luego pulse en el botón Aceptar (OK) para que se ejecute la orden. Los resultados obtenidos aparecen en la tabla siguiente:

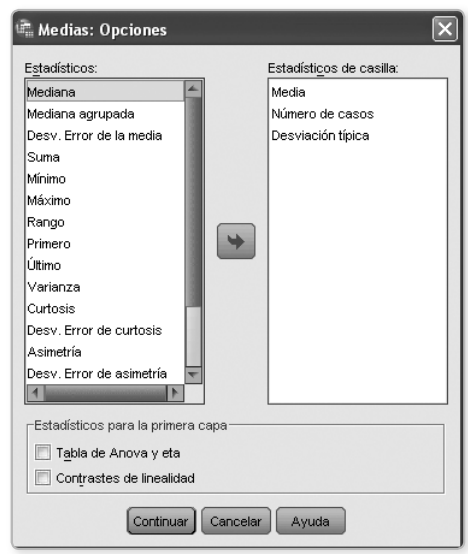
**Report**

Venta de Artículos de Aseo Personal

| Edad | Mean | N | Std. Deviation |
|---|---|---|---|
| Adolescente | 2.2667 | 3 | 2.1127 |
| Joven | 1.7200 | 5 | 1.1367 |
| Adulto | 2.9182 | 22 | 2.4510 |
| Total | 2.6533 | 30 | 2.2483 |

La tabla muestra una comparación de las medias de la venta de artículos de Aseo personal, en cada subgrupo de la variable Edad. De aquí, se obtiene que la media de los adolescentes en la compra de artículos de aseo personal es 2.2667 y la media de los jóvenes en la compra de estos mismos artículos es 1.7200.

Además, si pulsa el botón Opciones, se obtiene como resultado el cuadro de diálogo siguiente:



Donde se observa un listado de estadística como media, mediana, valor mínimo, valor máximo; es decir, todas las estadísticas que se aprecian en el cuadro izquierdo.

Para poder realizar los cálculos, simplemente se deben pasar las estadísticas elegidas al cuadro de la derecha (celda de estadísticas). Dichas estadísticas se trasladan de un cuadro a otro mediante el botón  . Para ello, hay que trasladar al cuadro de la derecha las estadísticas (mínimo, máximo) que se desean analizar del modo siguiente (puede trasladarse una o más estadísticas).

Se obtiene la siguiente tabla de resultados:

**Report**

Venta de Artículos de Aseo Personal

| Edad | Mean | N | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Adolescente | 2.2667 | 3 | 2.1127 | .30 | 4.50 |
| Joven | 1.7200 | 5 | 1.1367 | .30 | 3.30 |
| Adulto | 2.9182 | 22 | 2.4510 | .20 | 6.50 |
| Total | 2.6533 | 30 | 2.2483 | .20 | 6.50 |

En la tabla de resultados se observan los valores de la media, desviación estándar, el valor mínimo, valor máximo de los grupos de edades en las Ventas de artículos de aseo personal. También se pueden comparar medias como: Media aritmética, Media armónica y Media geométrica, tal como se observa en el siguiente cuadro de resultados:

**Report**

Venta de Abarrotes

| Estado Civil | Mean | Harmonic Mean | Geometric Mean |
|---|---|---|---|
| Soltero | 36.5200 | 36.4116 | 36.4652 |
| Casado | 34.1833 | 34.1282 | 34.1559 |
| Divorciado | 34.1500 | 34.0483 | 34.0993 |
| Viudo | 34.0889 | 33.8444 | 33.9674 |
| Total | 34.5400 | 34.3904 | 34.4655 |

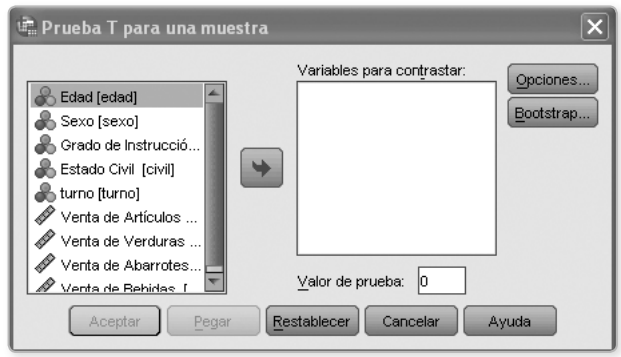
En la tabla de resultados, se puede observar la comparación de medias de los grupos de edades en las ventas de abarrotes.

### B. Prueba T

Para comparar medias utilizando la prueba T, hay tres opciones diferentes que permiten realizar contrastes de hipótesis sobre una media, dos medias con muestras independientes y dos medias con muestras relacionadas.

### C. Prueba T para una muestra

Es el contraste sobre una media. Al seleccionar la opción menú **Analizar > Comparar medias > Prueba T para una muestra** (one-sample T Test), se muestra el cuadro de diálogo siguiente:




El cuadro de la izquierda muestra el listado de las variables que posee el archivo de datos que está abierto. Los cuadros de la derecha están vacíos, pues esperan la especificación de las variables a analizar. Dichas variables se trasladan de un cuadro a otro mediante el botón .

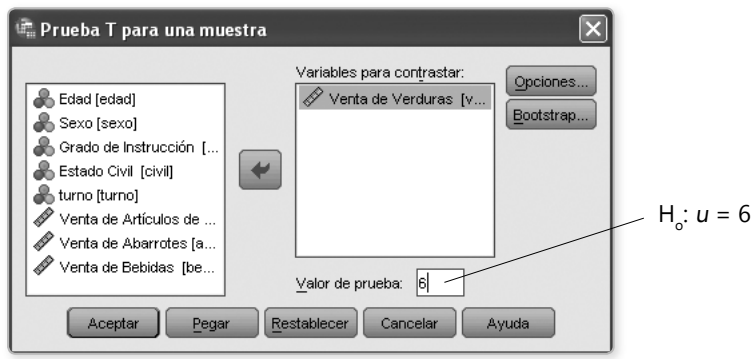
Para realizar el contraste, hay que trasladar al cuadro **Variables para contrastar**, Test Variable(s), la variable cuya media se desea contrastar o indicar en **Valor de prueba** (Test Value), el valor de la media poblacional en la hipótesis nula. Luego pulse el botón Aceptar (OK).

Por ejemplo, si se quiere comprobar si la media poblacional de la Venta de verduras es igual a 6, entonces la prueba de hipótesis será la siguiente:

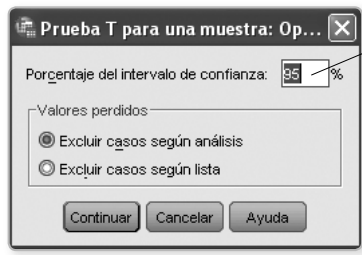
$$H_o: \mu = 6$$
$$H_1: \mu \neq 6$$

Con nivel de significancia de $\alpha = 0{,}05$. Es decir, con un nivel de confianza de $1-\alpha = 0{,}95$ (95%). En consecuencia, se rechaza $H_o$ si $\alpha >$ Sig y se acepta $H_1$. Pero si $\alpha \leq$ Sig, se acepta $H_0$ y se rechaza $H_1$.

Se seleccionan las opciones mostradas en cuadro de diálogo de la Prueba T para una muestra (One Sample T Test).



Además, si pulsa sobre el botón Opciones, puede modificarse el nivel de confianza que, por defecto, es 0,95 (95%). El cuadro de diálogo Opciones es el siguiente:




Nivel de confianza

El resultado del análisis se muestra en el Visor de resultados. SPSS proporciona dos tablas de resultados: en la primera aparece el número de casos, la media, la desviación típica y el error típico de la media.

$$\left(\sigma_{\bar{x}} = S_{n-1} / \sqrt{n}\right)$$

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| Venta de Verduras | 30 | 12.5333 | 5.7219 | 1.0447 |

En la segunda tabla se muestran los resultados de la prueba T: el valor de la estadística de contraste, los grados de libertad (df), el nivel crítico bilateral, la diferencia entre la media muestral y la media poblacional (μ) de la hipótesis nula y, por último, los límites (inferior y superior) del intervalo de confianza para la diferencia.

($\underline{x}$ - μ; donde $\underline{x}$ es la media muestral).

Estadístico de contraste

Nivel crítico

**One-Sample Test**

| | Test Value = 6 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| Venta de Verduras | 6.254 | 29 | .000 | 6.5333 | 4.3967 | 8.6699 |

Cuando se realizan contrastes de hipótesis con el SPSS, la decisión sobre $H_0$ debe tomarse a partir del nivel crítico, que es la probabilidad asociada al estadístico de contraste.

SPSS suele ofrecer el nivel crítico bilateral. En el ejemplo, el valor de T ha resultado ser 6,254 y el nivel crítico bilateral (Sig) 0,000. Esto significa que P (T ≤ -6.254) + P (T ≥ 6.254) = 0,000 lo que le permite rechazar $H_0$ con $\alpha$ = 0,05 ($\alpha$ = 0,05 > Sig = 0.000). Esta decisión también puede tomarse a partir del intervalo de confianza para la diferencia de medias, el cual no incluye el valor 0.

Para poder visualizar mejor se presenta el gráfico de la **distribución T bilateral** (dos colas).

Hipótesis planteada:

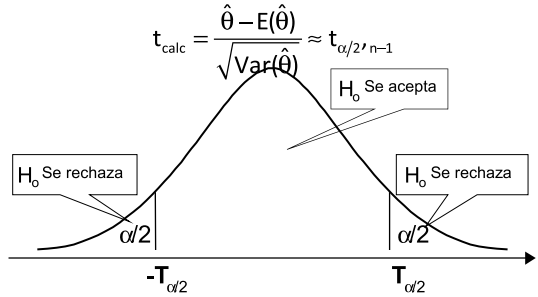
$H_o$: μ = μ o

$H_1$: μ ≠ μ o

Nivel de significancia: $\alpha$%

Estadístico de prueba:

$$t_{calc} = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{Var(\hat{\theta})}} \approx t_{\alpha/2,\, n-1}$$




Regla de decisión:

$T < -T_{\alpha/2}$ y $T > T_{\alpha/2}$: Se rechaza la hipótesis $H_o$.

$-T_{\alpha/2} < T < T_{\alpha/2}$: Se acepta la hipótesis $H_o$.

En caso de haber deseado realizar un contraste unilateral, es necesario dividir por dos el nivel crítico bilateral. Continuando con el ejemplo, para contrastar la hipótesis:

$$H_0: \mu \leq 6$$

$$H_1: \mu > 6$$

Hay que dividir por dos el nivel crítico bilateral: 0,000/2 = 0,00 que es la probabilidad correspondiente al suceso: $T \geq 6.264$. En este caso, también se rechazará $H_0$ con $\alpha = 0,05$.

### D. Prueba T para muestras independientes

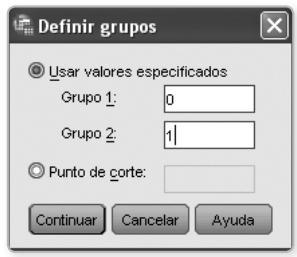
Es el contraste sobre dos medias independientes. Al seleccionar la opción menú Analizar > Comparar medias > Prueba T para muestras independientes (Idependent–Samples T Test) aparece el cuadro de diálogo siguiente:




Trasladar aquí la variable sobre la que se desea realizar el contraste.

Para realizar el contraste, es necesario indicar cómo están definidos los dos grupos en el Editor de datos.

En el cuadro de diálogo se ha indicado un contraste de las medias de las variables Venta de verduras, en los dos grupos de la variable Sexo. Además, es necesario indicar cuál es el valor de la variable Sexo en cada grupo, para lo cual se pulsa el botón Definir grupos (Define Groups), con lo que aparece un subcuadro de diálogo.




De este modo se indica que se van a comparar los dos grupos, en los que Sexo toma valores 0 para mujer y 1 para varón.

La prueba T sobre dos medias independientes tiene dos versiones, dependiendo de si se asumen varianzas iguales o distintas.

SPSS ejecuta la prueba de Levene para contrastar la hipótesis nula de que las varianzas poblacionales de los dos ***grupos son iguales***. En la siguiente línea, presenta la prueba de contrastar la hipótesis nula de que las varianzas poblacionales de los dos ***grupos son diferentes***.

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_1: \sigma^3_1 \neq \sigma^3_2$$

Con un nivel de significancia de $\alpha$; es decir, con un nivel de confianza de 1-$\alpha$, en consecuencia, se rechaza $H_o$ si $\alpha >$ Sig y se acepta $H_1$; sin embargo, si $\sigma \leq$ Sig, se acepta $H_0$ y se rechaza $H_1$.

A continuación, se proporciona el resultado de las dos versiones de la prueba T.

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| Venta de Verduras | Equal variances assumed | .262 | .613 | -1.218 | 28 | .233 | -3.7115 | 3.0479 | -9.9549 | 2.5318 |
| | Equal variances not assumed | | | -.996 | 3.564 | .382 | -3.7115 | 3.7258 | -14.5744 | 7.1513 |

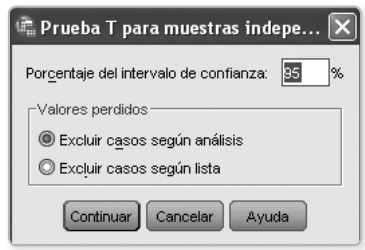
Es responsabilidad del usuario escoger el resultado de una versión de la prueba T, en función del resultado del contraste sobre varianzas. El nivel crítico de cada contraste aparece en la columna Sig.

La hipótesis nula de la prueba de Levene es $H_0: \sigma^2_1 = \sigma^2_2$, es decir, que la varianza poblacional es igual en ambos grupos. En este ejemplo, el estadístico de contraste (con valor F = 0,262) no ha resultado significativo, según indica el nivel crítico igual a 0,613; por tanto, al mantenerse $H_0$, puede asumirse que las varianzas son iguales.

Se puede observar que en la tabla de resultados aparecen dos valores de T. En este caso, se utiliza el primero de ellos correspondiente al supuesto de igualdad de varianzas (T = -1.218).

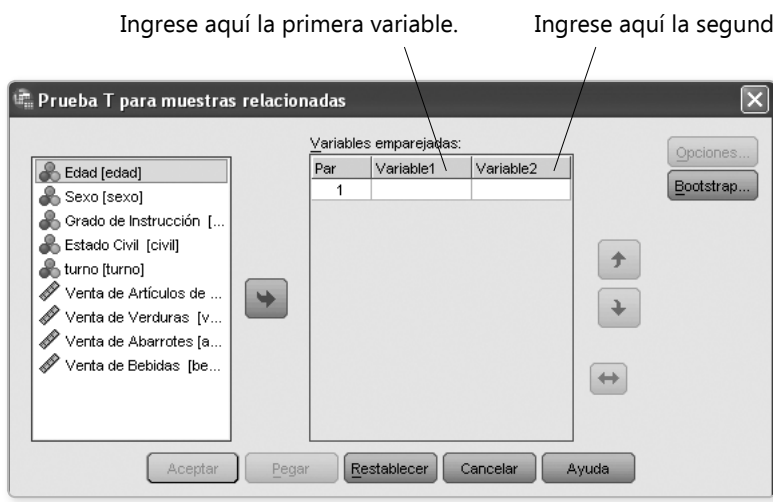
De la misma forma, la decisión sobre la hipótesis nula debe tomarse a partir del nivel crítico (igual a 0,233) que, en este caso, indica que el valor de T no es significativo y, por tanto, que las medias de la Venta de verduras son iguales en varones y mujeres.

Además, si pulsa el botón Opciones puede modificar el nivel de confianza que, por defecto, es 0,95.




### E. Prueba T para muestras relacionadas

Es el contraste sobre dos medias relacionadas. Al seleccionar la opción menú Analizar > Comparar medias > Prueba T para muestras relacionadas (Paired Samples T Test) aparece el cuadro de diálogo siguiente.



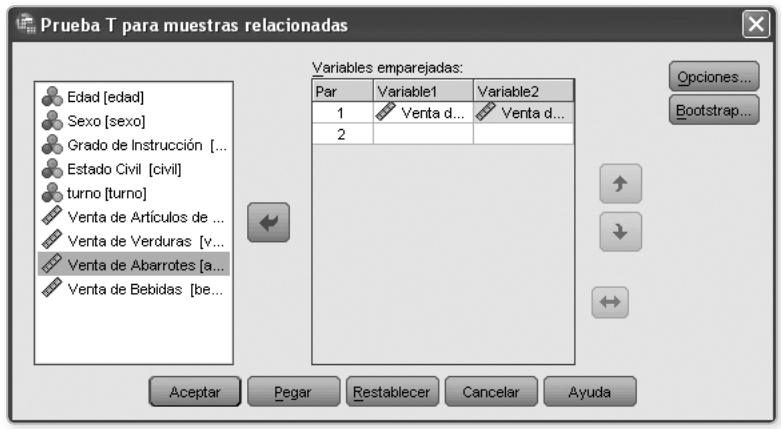

En la siguiente línea se presenta la prueba de contrastar la hipótesis nula de que las varianzas poblacionales de los dos grupos son diferentes.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Con un nivel de significancia de $\alpha$, es decir con un nivel de confianza de $1-\alpha^2$. En consecuencia, se rechaza $H_o$ si $\alpha >$ Sig y se acepta $H_1$. Pero si $\alpha \leq$ Sig, se acepta $H_0$ y se rechaza $H_1$.

Para comparar las medias con dos muestras relacionadas es necesario seleccionar dos variables de la lista, tal y como se observa en el cuadro de diálogo. Luego, para contrastar la hipótesis nula de que la diferencias de medias de Venta de verduras (variable 1) y Venta de abarrotes (variable 2) es significativamente diferente de cero, se pulsa sobre el botón Flecha ( ) para trasladar las variables al cuadro de variables seleccionadas, tal como se muestra en la figura siguiente:




Luego, pulse el botón Aceptar (OK) para que se ejecute la orden contraste. Los resultados obtenidos aparecen en la ventana Visor de resultados. Los resultados obtenidos son los siguientes:

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | Venta de Verduras | 12.5333 | 30 | 5.7219 | 1.0447 |
| | Venta de Abarrotes | 34.5400 | 30 | 2.2997 | .4199 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | Venta de Verduras & Venta de Abarrotes | 30 | -.187 | .323 |

**Paired Samples Test**

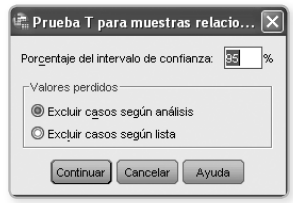
| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | Venta de Verduras - Venta de Abarrotes | -22.01 | 6.5535 | 1.1965 | -24.45 | -19.56 | -18.4 | 29 | .000 |

Para realizar este contraste, se calcula una nueva variable que es la diferencia entre las originales:

D = Venta de verduras – Venta de abarrotes

La salida de resultados muestra la media de la variable D, así como su desviación típica, el error típico de la media y el intervalo de confianza. También puede verse el estadístico de contraste T, los grados de libertad

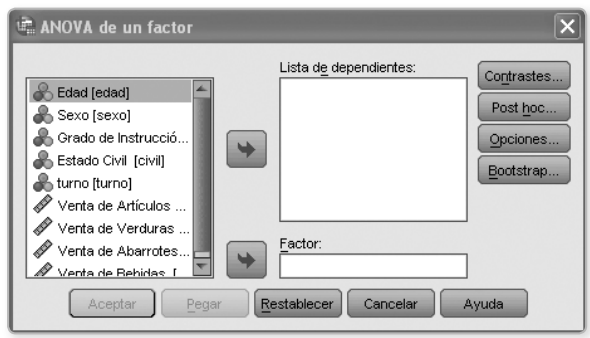
(df) y el nivel crítico bilateral. Puesto que el nivel crítico vale p = 0,000 se rechaza $H_0$ con un nivel de confianza de 0,95 y se concluye que la media en la población de diferencia, es significativamente distinta de cero. Es decir, existen diferencias significativas entre el promedio de ventas tanto de Verduras como de Abarrotes.
Además, si se pulsa el botón Opciones puede modificarse el nivel de confianza que, por defecto, es 0,95. El cuadro de diálogo Opciones es el siguiente:



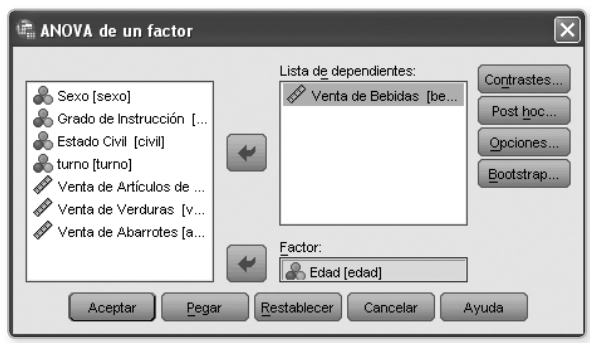

### F. ANOVA de un factor

Es el análisis de varianza de un factor en un diseño completamente aleatorizado ANOVA A-EF-CA. Existen distintos procedimientos en SPSS para realizar un análisis de varianza de un factor en un diseño completamente aleatorizado.

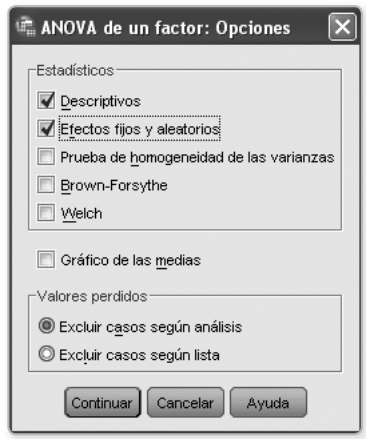
Uno de los más directos y sencillos se encuentra dentro del menú **Analizar** > **Comparar medias** > **ANOVA de un factor** (One-Way ANOVA). Se obtiene el siguiente cuadro de diálogo:




Por ejemplo, para contrastar la hipótesis de igualdad de medias en la variable venta de bebidas, en los tres grupos de Edad (Adolescente, Joven y Adulto), se ingresa al menú Analizar > Comparar medias > ANOVA de un factor y se procede a llenar el cuadro de diálogo de la manera siguiente:



Dentro del cuadro de diálogo ANOVA de un factor hay varios submenús. Por ejemplo, si pulsa el botón Opciones puede indicarse que SPSS calcule estadísticos descriptivos para las variables en los "*n*" grupos, y también que compruebe el supuesto de «homocedasticidad» (homogeneidad de varianzas). El cuadro de diálogo Opciones es el siguiente:




Si se pulsa Aceptar (OK), el resultado obtenido es el siguiente:

**Descriptives**

Venta de Bebidas

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| Adolescente | 3 | 5.0000 | 1.7321 | 1.0000 | .6973 | 9.3027 | 3.00 | 6.00 |
| Joven | 5 | 4.4000 | 2.3022 | 1.0296 | 1.5415 | 7.2585 | 1.00 | 6.00 |
| Adulto | 22 | 4.0000 | 1.6903 | .3604 | 3.2506 | 4.7494 | 1.00 | 6.00 |
| Total | 30 | 4.1667 | 1.7633 | .3219 | 3.5082 | 4.8251 | 1.00 | 6.00 |

Lo primero que aparece en el Visor de resultados son los estadísticos descriptivos para los "n" (3) grupos. Como se observa, el grupo que más gasta en bebidas son los Adolescentes, y el que menos gasta en esto son los Adultos.

A continuación, se muestra el resultado del test de Levene sobre homogeneidad de varianzas, cuya hipótesis nula es como sigue:

$$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$$

$$H_1: \sigma^2_1 \neq \sigma^2_2 \neq \sigma^2_3$$

Con un nivel de significancia de $\alpha$ ; es decir, con un nivel de confianza del 1-$\alpha$. Por tanto, se rechaza $H_o$ si $\alpha$ > Sig y se acepta $H_1$; pero si $\alpha \leq$ Sig, se acepta $H_0$ y se rechaza $H_1$.

**Test of Homogeneity of Variances**

Venta de Bebidas

| Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|
| .798 | 2 | 27 | .460 |

Puesto que el nivel crítico (p = 0,460) es mayor que $\alpha$, se mantiene la hipótesis nula sobre igualdad de varianzas; por tanto, se asume que se cumple el supuesto de homocedasticidad (homogeneidad de varianzas). A continuación, SPSS muestra la tabla resumen ANOVA:

**ANOVA**

Venta de Bebidas

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 2.967 | 2 | 1.483 | .459 | .637 |
| Within Groups | 87.200 | 27 | 3.230 | | |
| Total | 90.167 | 29 | | | |

Seguidamente, se muestra el contraste de la prueba F, cuya hipótesis nula es la siguiente:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

Con un nivel de significancia de $\alpha$ ; es decir, con un nivel de confianza del 1-$\alpha$. Por tanto, se rechaza $H_o$ si $\alpha$ > Sig y se acepta $H_1$; pero si $\alpha \leq$ Sig, se acepta $H_0$ y se rechaza $H_1$.

En la tabla se muestran las sumas de cuadrados intergrupos, intragrupos (error) y total. La tabla también contiene los grados de libertad, medias cuadráticas, estadísticos de contraste F y su nivel crítico (Sig.). En este ejemplo, puesto que el nivel crítico (0.6379) es mayor a 0,05, se acepta la hipótesis nula del ANOVA ($H_0$: $\mu_1 = \mu_2 = \mu_3$), y puede concluirse que no existen diferencias significativas en el gasto medio de bebidas por los sujetos de cada edad (adolescente, joven y adulto).

## G. Comparaciones múltiples entre medias

Son las comparaciones de tendencia y F planeadas. Para realizar dichas comparaciones es necesario pulsar el botón Contrastes, que se encuentra situado en el cuadro de diálogo del procedimiento ANOVA de un factor.

Como se ve el siguiente ejemplo:

Si se desea realizar la comparación de tendencia sobre el componente cuadrático, se indicará del siguiente modo:




número de media es *n = 3*

SPSS proporciona los resultados del componente indicado y también de los de orden inferior; en este ejemplo, puesto que *n = 3*, constituye el cuadrático y el lineal.

**ANOVA**

Venta de Bebidas

| | | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|---|
| Between Groups | (Combined) | | 2.967 | 2 | 1.483 | .459 | .637 |
| | Linear Term | Unweighted | 2.640 | 1 | 2.640 | .817 | .374 |
| | | Weighted | 2.933 | 1 | 2.933 | .908 | .349 |
| | | Deviation | 3.393E-02 | 1 | 3.393E-02 | .011 | .919 |
| | Quadratic Term | Unweighted | 3.393E-02 | 1 | 3.393E-02 | .011 | .919 |
| | | Weighted | 3.393E-02 | 1 | 3.393E-02 | .011 | .919 |
| Within Groups | | | 87.200 | 27 | 3.230 | | |
| Total | | | 90.167 | 29 | | | |

La tabla de resultados proporciona la F del ANOVA. Puede verse que su valor es 0.459 en este ejemplo (el mismo que en la salida de resultados comentada anteriormente). Además, proporciona la F del componente lineal en la fila denominada «no ponderado», que toma el valor 0.817, y la del componente cuadrático, que ha resultado ser 0.011. La columna Sig se refiere a la hipótesis sobre cada tendencia, que se acepta en ambos casos con $p < 0.05$ en la tendencia lineal, y $p < 0.01$, en la cuadrática. Por tanto, la relación entre ambas variables es lineal.

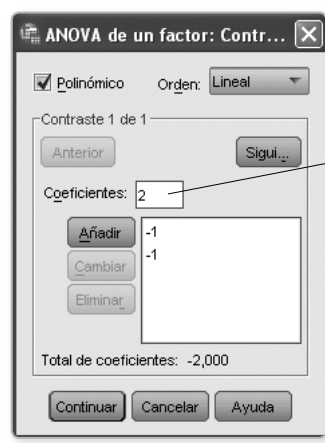
Para llevar a cabo ***comparaciones planeadas*** es necesario introducir los coeficientes directamente. Por ejemplo, para comparar la media en venta de bebidas del grupo 3 (adulto) de Edad, con la media de los grupos 1 (adolescente) y 2 (joven) tomados juntos, la hipótesis planteada sería la siguiente:

$$H_0: (-1)\,\mu_1 + (-1)\,\mu_2 + (2)\,\mu_3 = 0$$

$$H_1: (-1)\,\mu_1 + (-1)\,\mu_2 + (2)\mu_3 \neq 0$$

Con un nivel de significancia de $\alpha$; es decir, con un nivel de confianza de $1-\alpha$. En consecuencia, se rechaza $H_o$ si $\alpha >$ Sig y se acepta $H_1$; sin embargo, si $\alpha \leq$ Sig, se acepta $H_0$ y se rechaza $H_1$.

En SPSS, esto se indica mediante:




Escriba aquí el valor de cada coeficiente y pulse Añadir.

La salida de los resultados incluye los coeficientes que se han utilizado en la comparación, lo cual permite comprobar si han sido asignados correctamente a los grupos. También incluye el estadístico T del contraste de medias, en sus dos versiones: asumiendo varianzas iguales y distintas.

Puesto que la prueba de Levene indicó igualdad de varianzas, el estadístico de prueba T vale -0.921, y su nivel crítico 0,365; por tanto, se mantiene $H_0$ y no puede concluirse que existan diferencias significativas entre los sujetos del grupo adolescentes y jóvenes frente a los adultos, tal como se aprecia en los siguientes cuadros:

**Contrast Coefficients**

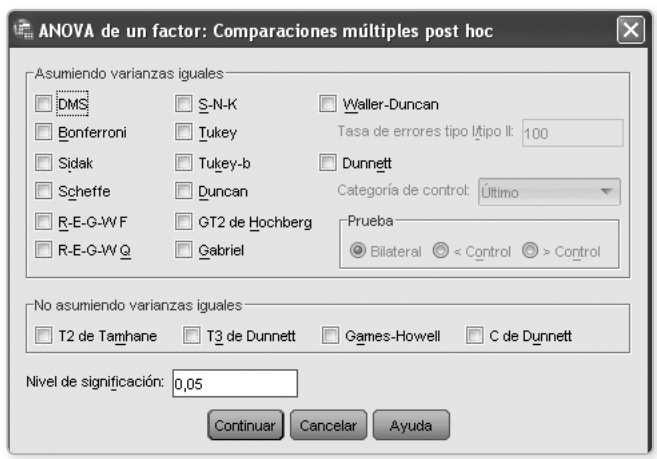
| | Edad | | |
|---|---|---|---|
| Contrast | Adolescente | Joven | Adulto |
| 1 | -1 | -1 | 2 |

**Contrast Tests**

| | | Contrast | Value of Contrast | Std. Error | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|
| Venta de Bebidas | Assume equal variances | 1 | -1.4000 | 1.5198 | -.921 | 27 | .365 |
| | Does not assume equal | 1 | -1.4000 | 1.6061 | -.872 | 8.383 | .408 |

### H. Comparaciones a posteriori

Al pulsar el botón Post Hoc, se muestra el cuadro de diálogo correspondiente a las comparaciones a posteriori:




En el cuadro de diálogo, se puede ver que permite seleccionar distintas pruebas como: Tukey, Dunnett, Scheffé, etc. Continuando con el ejemplo, al seleccionar la prueba de Tukey, la salida SPSS muestra la tabla de resultados de dicha prueba.

En la tabla puede verse la diferencia entre las medias de la venta de ***Bebidas*** en cada uno de los pares de grupos definidos por la variable ***Edad***. También se incluye el error típico de la diferencia de medias, el nivel crítico asociado a dicha diferencia (Sig.) y los límites del intervalo de confianza. Como se observa, los resultados indican que no existen diferencias significativas entre el grupo adolescentes y joven ($p < 0,05$) y adolescentes y adulto ($p < 0,05$), o el grupo joven y adolescente ($p < 0,05$) y joven y adulto ($p < 0,05$).

**Multiple Comparisons**

Dependent Variable: Venta de Bebidas
Tukey HSD

| (I) Edad | (J) Edad | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Adolescente | Joven | .6000 | 1.3124 | .892 | -2.6541 | 3.8541 |
| | Adulto | 1.0000 | 1.1060 | .642 | -1.7424 | 3.7424 |
| Joven | Adolescente | -.6000 | 1.3124 | .892 | -3.8541 | 2.6541 |
| | Adulto | .4000 | .8904 | .895 | -1.8076 | 2.6076 |
| Adulto | Adolescente | -1.0000 | 1.1060 | .642 | -3.7424 | 1.7424 |
| | Joven | -.4000 | .8904 | .895 | -2.6076 | 1.8076 |

Grupos

### 5.1.2. EJERCICIOS PROPUESTOS

1. Los resultados obtenidos por un gabinete psicológico durante los últimos 15 años, indican que los pacientes con insomnio duermen un promedio de 7 horas durante la tercera noche. Contraste dicha afirmación utilizando =0,05.

2. Se desea saber si los sujetos que han recibido la terapia 2, duermen durante la segunda noche el mismo número de horas que los que la han recibido. Utilice = 0,01.

3. Los datos representan 10 sujetos que participaron en el estudio, cuyo resultado se muestra en la tabla.

   Los datos aparecen recopilados en las variables Día1, Día2, Día3 que indican el número de horas trabajadas en los tres días. Además, se ha recopilado la variable ***Sexo,*** que toma valores 0 para mujeres y 1 para hombres. ***Edad***, 1 para adolescentes, 2 para jóvenes y 3 para adultos. Finalmente, se incluyen tres datos relacionados con las terapias que ha recibido el sujeto. Todas ellas toman 1, si el sujeto ha recibido la terapia y 0 en caso contrario. Las variables son ***Estrés***, ***Ansiedad*** y ***Fobia***.

| Sujeto | Día1 | Día2 | Día3 | Sexo | Edad | Estrés | Ansiedad | Fobia |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.7 | 6.3 | 6.8 | 0 | 1 | 0 | 1 | 0 |
| 2 | 4.8 | 5.4 | 4.3 | 1 | 3 | 0 | 0 | 0 |
| 3 | 7.6 | 7.9 | 8.6 | 0 | 2 | 0 | 0 | 0 |
| 4 | 5.7 | 6.0 | 6.7 | 1 | 1 | 0 | 0 | 1 |
| 5 | 3.8 | 4.6 | 4.9 | 1 | 3 | 0 | 0 | 0 |
| 6 | 7.5 | 8.0 | 8.5 | 1 | 2 | 0 | 1 | 0 |
| 7 | 6.4 | 7.0 | 7.3 | 1 | 1 | 1 | 1 | 0 |
| 8 | 7.7 | 8.1 | 8.5 | 0 | 2 | 0 | 0 | 0 |
| 9 | 4.0 | 4.5 | 5.0 | 0 | 3 | 1 | 0 | 0 |
| 10 | 5.7 | 6.1 | 6.8 | 1 | 2 | 1 | 0 | 1 |

   Se desea saber si los sujetos trabajan más horas durante el tercera día que durante el primero (con = 0,05).

4. Del ejercicio propuesto 3, contraste la hipótesis de que la media en horas trabajadas durante el tercer día, para la población que recibió terapia con estrés, es de 8 horas (= 0,99).

5. Una escuela de secundaria ha contratado a un psicólogo clínico, buscando que invente una terapia que sirva para reducir el nivel de ansiedad de sus alumnos ante los exámenes. Para ello, selecciona aleatoriamente a ocho alumnos y mide su nivel de ansiedad en los exámenes de febrero, con una escala de 1 a 10 puntos. A continuación, les aplica la terapia y vuelve a medir su nivel de ansiedad en los exámenes de junio. Los resultados aparecen en la tabla inferior. Según esto, ¿qué concluirá el psicólogo sobre la eficacia de su terapia con = 0,01?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Antes | 10 | 8 | 9 | 7 | 8 | 5 | 9 | 7 |
| Después | 6 | 5 | 7 | 6 | 4 | 5 | 4 | 6 |

6. Con los datos del ejercicio 3, un psicólogo desea comprobar si existen diferencias en el número medio de horas trabajadas por los pacientes con estrés, durante la primera noche, en los tres grupos de edad considerados.
   a. Seleccione el modelo de ANOVA más apropiado, plantee la hipótesis pertinente y tome una decisión con =0,05.
   b. ¿Existe relación entre la edad y el número de horas trabajadas el primer día? En caso afirmativo, ¿cuál es la tendencia de dicha relación?
   c. ¿A qué edad los pacientes con estrés trabajan menos horas? (=0,05).
   d. Compruebe si el número de horas de trabajo de los sujetos jóvenes difieren del nivel de los restantes considerados juntos (= 0,01).

7. Repita los apartados a, b, c y d del problema anterior; pero referido primero al segunda día y después al tercer día.

8. Un psicólogo clínico desea evaluar la eficacia de un fármaco para reducir la ansiedad. Para ello, selecciona al azar 15 pacientes de su consulta que sufren este problema, y forma aleatoriamente tres grupos del mismo tamaño. A cada grupo le administra aleatoriamente una dosis del fármaco (10, 20 y 30 miligramos).

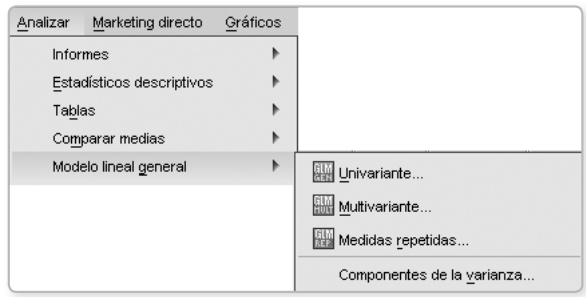
   Al cabo de un tiempo, le mide su nivel de ansiedad. Los resultados obtenidos se muestran en la siguiente tabla:

| | | | | | |
|---|---|---|---|---|---|
| 10 mg | 7 | 8 | 8 | 9 | 8 |
| 20 mg | 4 | 4 | 5 | 6 | 6 |
| 30 mg | 2 | 3 | 2 | 2 | 1 |

   a. Indique las variables que intervienen en el problema y la función que desempeñan.
   b. Seleccione el modelo de ANOVA más apropiado, plantee la hipótesis pertinente y tome una decisión con = 0,05.
   c. ¿Existe relación entre la dosis del fármaco y el nivel de ansiedad?
   d. En caso afirmativo, interpreta gráfica y estadísticamente el tipo de relación.
   e. Cuantifica el tamaño de dicha relación.
   f. ¿Entre qué dosis existen diferencias?
   g. Compruebe si el nivel de ansiedad con la dosis de 10 mg difiere del nivel de las restantes, consideradas juntas.

## 5.2. MODELO LINEAL GENERAL

El modelo lineal general permite llevar a cabo ANOVAS de un factor o cualquier otro tipo de ANOVA. Los procedimientos se muestran a continuación:




### 5.2.1. CONCEPTOS BÁSICOS

#### A. Univariante

Proporciona un análisis de regresión y un análisis de varianza para una variable dependiente, mediante uno o más factores y variables. Las variables de factor dividen la población en grupos. Con este procedimiento, se pueden contrastar hipótesis nulas sobre los efectos de otras variables, en las medias de varias agrupaciones de una única variable dependiente. Se pueden investigar las interacciones entre los factores, así como los efectos de los factores individuales, algunos de los cuales pueden ser aleatorios. Además, se pueden incluir los efectos de las covariables y las interacciones de covariables con los factores.

Para el análisis de regresión, las variables independientes (predictoras) se especifican como covariables. Se pueden contrastar tanto los modelos equilibrados, como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos; además de contrastar las hipótesis, se generan estimaciones de los parámetros.

También se encuentran disponibles los contrastes de hipótesis de uso más habitual. Si una prueba F global ha mostrado cierta significación, pueden emplearse las pruebas post hoc para evaluar las diferencias entre las medias específicas. Las medias marginales estimadas ofrecen estimaciones de valores de las medias, pronosticados para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones.

En el archivo de datos se pueden guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos.

- **Ponderación MCP.** Permite especificar una variable usada, para aplicar a las observaciones una ponderación diferente, en un análisis de mínimos cuadrados ponderados (MCP). Por ejemplo, para compensar la distinta precisión de las medidas.
- **Métodos.** Las sumas de cuadrados de Tipo I, Tipo II, Tipo III y Tipo IV pueden emplearse para evaluar las diferentes hipótesis. El Tipo III es el valor por defecto.
- **Gráficos.** Diagramas de dispersión por nivel, gráficos de residuos y gráficos de perfil (interacción).

### B. Multivariante

Proporciona un análisis de regresión y un análisis de varianza, para variables dependientes múltiples por una o más covariables, o variables de factor. Las variables de factor dividen la población en grupos. Con este procedimiento, es posible contrastar hipótesis nulas sobre los efectos de las variables de factor, sobre las medias de varias agrupaciones de una distribución conjunta, de variables dependientes.

En un modelo multivariado, las sumas de cuadrados debidas a los efectos del modelo y las sumas de cuadrados error, se encuentran en forma de matriz, en lugar de la forma escalar del análisis univariado. Estas matrices se denominan «matrices SCPC» (sumas de cuadrados y productos cruzados). Si se especifica más de una variable dependiente, se proporciona el análisis multivariado de varianzas, usando la traza de Pillai, la lambda de Wilks, la traza de Hotelling y el criterio de mayor raíz de Roy, con el estadístico F aproximado, así como el análisis univariado de varianza para cada variable dependiente. Las pruebas de comparaciones múltiples post hoc se realizan por separado para cada variable dependiente.

También se hallan disponibles: una **matriz SCPC residual,** que es una matriz cuadrada de las sumas de cuadrados y los productos cruzados de los residuos; una **matriz de covarianza residual,** que es la matriz SCPC residual dividida por los grados de libertad de los residuos; y la **matriz de correlaciones residual,** que es la forma tipificada de la matriz de covarianza residual.

El resto es similar al procedimiento anterior.

### C. Medidas repetidas

Analiza grupos de variables dependientes relacionadas, que representan diferentes medidas del mismo atributo. Permite definir uno o varios factores intrasujetos (no confundir con las variables existentes), donde cada factor constituye un nivel dentro del factor precedente. Si los sujetos se comparan en más de una medida cada vez, seleccione Medida para definirlas. El nombre de las medidas no existe como un nombre de variable en el propio archivo de datos, sino que se define aquí.

### D. Componentes de la varianza

Estima la contribución de cada efecto aleatorio para modelos de efectos mixtos, a la varianza de la variable dependiente. Este procedimiento resulta de interés para el análisis de modelos mixtos, como los diseños split-plot, los diseños de medidas repetidas univariados y los diseños de bloques aleatorios. Al calcular los componentes de la varianza, se puede determinar dónde centrar la atención para reducirla. Se dispone de cuatro métodos diferentes para estimar las componentes de la varianza:

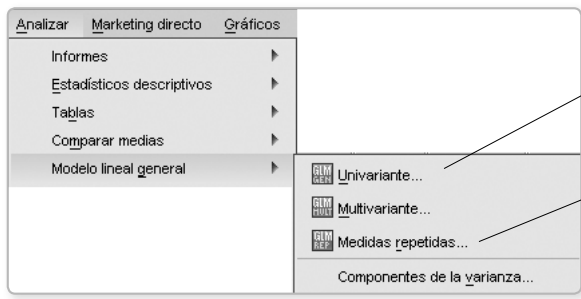
- Estimador mínimo no cuadrático insesgado (EMNCI, MINQUE)
- Análisis de varianza (ANOVA)
- Máxima verosimilitud (MV, ML)
- Máxima verosimilitud restringida (MVR, RML)

Además, existen diversas especificaciones para los diferentes métodos. Los resultados por defecto para todos los métodos, incluyen las estimaciones de componentes de la varianza. Si se usa el método MV o el método MVR, se mostrará también una tabla con la matriz de covarianza asintótica. Otros resultados disponibles incluyen una tabla de ANOVA, las medias cuadráticas esperadas para el método ANOVA y la historia de iteraciones para los métodos MV y MVR. Este procedimiento es totalmente compatible con el procedimiento MLG (modelo lineal general) factorial general.

### 5.2.2. EJERCICIO PRÁCTICO 1

El ejercicio práctico permitirá utilizar los principales procedimientos del menú Modelo lineal general. Los datos usados serán los recopilados en la encuesta del supermercado.

En el tema anterior, se ha visto el menú Comparar medias para llevar a cabo ANOVAS de un factor. En SPSS hay otro menú que puede utilizarse para llevar a cabo cualquier tipo de ANOVA. Se trata del menú Analizar > Modelo lineal general, cuyo aspecto es el siguiente:




A continuación, se presenta una explicación detallada del cuadro de diálogo correspondiente a los principales procedimientos. Utilizando el procedimiento univariante, puede llevarse a cabo un ANOVA A-EF-CA. A continuación se verá este menú detalladamente en el tema ANOVA de dos factores, en un diseño completamente aleatorizado.

#### A. Análisis de varianza de dos factores

El software estadístico SPSS ofrece buenas posibilidades para realizar análisis de varianza en diseños factoriales, con efectos fijos, aleatorios y mixtos. El procedimiento se encuentra ubicado en la opción del menú **Analizar > Modelo lineal general**.

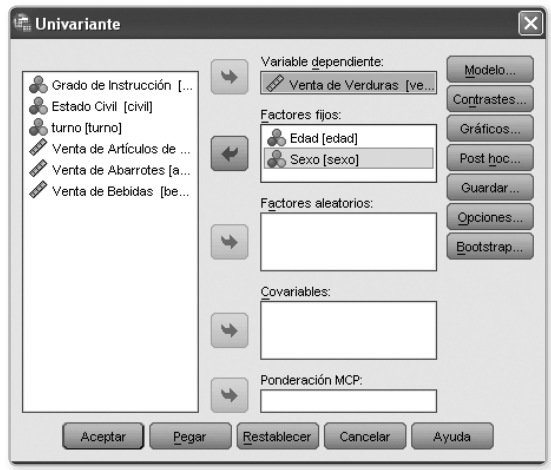
#### a. ANOVA AB–EF–CA

Este procedimiento estadístico permite realizar un análisis de varianza con más de un factor, incluyendo factores de efectos fijos y aleatorios, además de análisis de covarianza (ANCOVA).

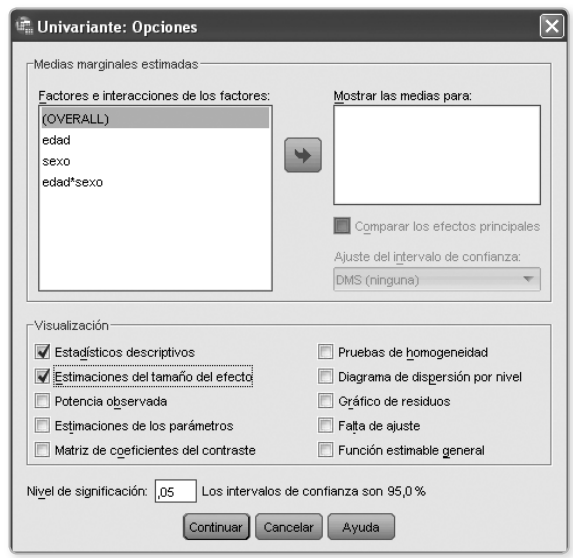
Suponiendo que desea realizar un ANOVA factorial de la variable *Venta de verduras*, utilizando como variable independiente el sexo del sujeto y su edad. Esto se especificaría del siguiente modo:

Ingrese al menú **Analizar > Modelo lineal general > Univariante** (univariate).

El procedimiento **Univariante** presenta el cuadro de diálogo siguiente:




El cuadro de diálogo del botón Opciones, del procedimiento Univariante, es el siguiente:




Pulsando Continuar y luego pulsando el botón Aceptar, SPSS realiza el ANOVA indicado y ofrece los resultados:

**Descriptive Statistics**

Dependent Variable: Venta de Verduras

| Edad | Sexo | Mean | Std. Deviation | N |
|---|---|---|---|---|
| Adolescente | Mujer | 20.0000 | 3.6056 | 3 |
| | Total | 20.0000 | 3.6056 | 3 |
| Joven | Mujer | 11.5000 | 2.6458 | 4 |
| | Varón | 24.0000 | . | 1 |
| | Total | 14.0000 | 6.0415 | 5 |
| Adulto | Mujer | 10.8947 | 5.1950 | 19 |
| | Varón | 13.0000 | 5.5678 | 3 |
| | Total | 11.1818 | 5.1606 | 22 |
| Total | Mujer | 12.0385 | 5.4734 | 26 |
| | Varón | 15.7500 | 7.1356 | 4 |
| | Total | 12.5333 | 5.7219 | 30 |

En Opciones, se han solicitado los estadísticos descriptivos, la prueba de homogeneidad y las estimaciones de tamaño del efecto.

El SPSS ofrece las medias y desviaciones típicas para las JK combinaciones de variables. Como se observa, quienes gastan más son los varones jóvenes.

**Levene's Test of Equality of Error Variances**

Dependent Variable: Venta de Verduras

| F | df1 | df2 | Sig. |
|---|---|---|---|
| .999 | 4 | 25 | .427 |

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+EDAD+SEXO+EDAD * SEXO

A continuación, se presenta la prueba de homocedasticidad, igualdad de varianzas, supuesta se cumple al mantener $H_0$.

La salida de resultados incluye la tabla –resumen del ANOVA y las medidas del tamaño del efecto:

**Tests of Between-Subjects Effects**

Dependent Variable: Venta de Verduras

| | Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| | Corrected Model | 354.677[a] | 4 | 88.669 | 3.727 | .016 | .374 |
| | Intercept | 2883.366 | 1 | 2883.366 | 121.193 | .000 | .829 |
| A | EDAD | 306.376 | 2 | 153.188 | 6.439 | .006 | .340 |
| B | SEXO | 130.390 | 1 | 130.390 | 5.481 | .028 | .180 |
| AB | EDAD * SEXO | 66.047 | 1 | 66.047 | 2.776 | .108 | .100 |
| E | Error | 594.789 | 25 | 23.792 | | | |
| T | Total | 5662.000 | 30 | | | | |
| | Corrected Total | 949.467 | 29 | | | | |

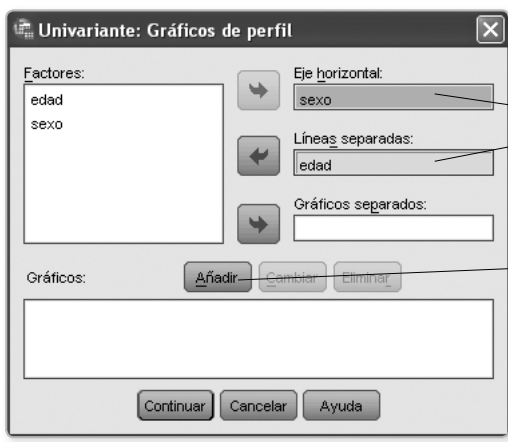
a. R Squared = .374 (Adjusted R Squared = .273)

$$\sum_i \sum_j \sum_k Y_{ijk}^2$$

Las estadísticas de contraste F correspondientes a los efectos de las variables *Edad* y *Sexo*, así como a la interacción entre ambas (Edad*Sexo), aparecen en la fila etiquetadas con el nombre de la variable. La decisión sobre $H_0$ se toma a partir del punto crítico. Puede verse que se rechaza la hipótesis nula correspondiente a la variable Edad, también se rechaza la variable Sexo y se mantiene la correspondiente a la interacción Edad* Sexo.

Para realizar un gráfico con los efectos de los factores, pulse el botón Gráficos del cuadro de diálogo del procedimiento univariante.

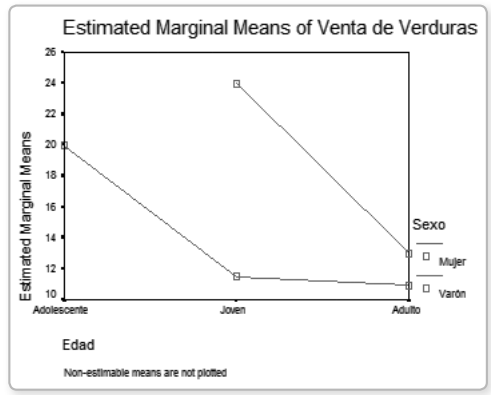
De este modo, un subcuadro de diálogo aparece en el que se especifica la forma donde se realizan uno o más gráficos. Por ejemplo:




Traslade la variable Sexo a la primera casilla y Edad a la segunda casilla.

Luego pulse Añadir (Add) gráfico a la lista.

De este modo, se ha especificado un gráfico de los valores de venta de verduras por Edad, con líneas distintas para cada nivel de Sexo. A continuación, pulse el botón Añadir para que este gráfico se incluya en la lista de gráficos que se van a realizar. La salida de resultados del ANOVA incluye todos los gráficos que se hayan especificado. En este caso:




El cuadro de diálogo del procedimiento Univariante incluye también el botón Post Hoc, que se utiliza para realizar contrastes a posteriori sobre los factores de efectos fijos. Para realizar la prueba de Tukey sobre los niveles de Edad, se indicará de la siguiente manera:




Los resultados de la prueba de Tukey aparecen junto con los de ANOVA. La tabla muestra las diferencias entre cada par de medias de la variable *Venta de verduras*, en los grupos de sujetos definidos por *Edad* y *Sexo*. Además, se incluye el error típico de cada diferencia, el nivel crítico y el intervalo de confianza para la diferencia de medias.

**Multiple Comparisons**

Dependent Variable: Venta de Verduras
Tukey HSD

| (I) Edad | (J) Edad | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Adolescente | Joven | 6.0000 | 3.5621 | .231 | -2.8727 | 14.8727 |
| | Adulto | 8.8182* | 3.0020 | .019 | 1.3407 | 16.2957 |
| Joven | Adolescente | -6.0000 | 3.5621 | .231 | -14.8727 | 2.8727 |
| | Adulto | 2.8182 | 2.4166 | .484 | -3.2011 | 8.8374 |
| Adulto | Adolescente | -8.8182* | 3.0020 | .019 | -16.2957 | -1.3407 |
| | Joven | -2.8182 | 2.4166 | .484 | -8.8374 | 3.2011 |

Based on observed means.
*. The mean difference is significant at the .05 level.

## B. Medidas repetidas

### a. Análisis de varianza de un factor con medidas repetidas

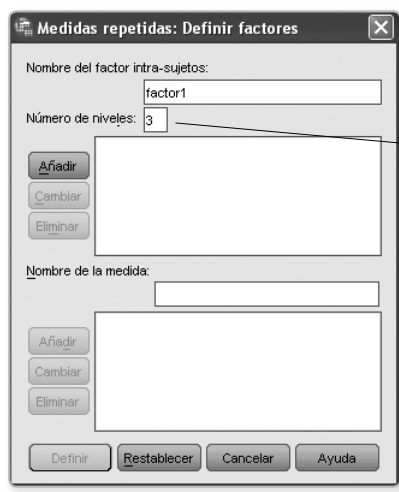
Anteriormente, se ha visto el menú Comparar medias para llevar a cabo ANOVAS de un factor. En SPSS, hay otro menú, que puede utilizarse para llevar a cabo cualquier tipo de ANOVA. Se trata del menú **Analizar > Modelo lineal general.**

### b. ANOVA A-EF-MR

Para realizar el ANOVA de un factor de diseño de medidas repetidas, se ejecuta la opción del menú **Analizar > Modelo lineal general > Medidas repetidas** (Repeated Measures, Define Factor[s]).

El objetivo es comparar las medias de distintas variables, medidas sobre los mismos sujetos. Por ejemplo, suponga que se desea comparar las medias de las *Venta de verduras*, *Venta de abarrotes* y *Venta de bebidas*, con el objeto de determinar la evolución de las ventas en cada tipos de artículos. Por tanto, se tiene un factor intrasujetos con $J = 3$ niveles.

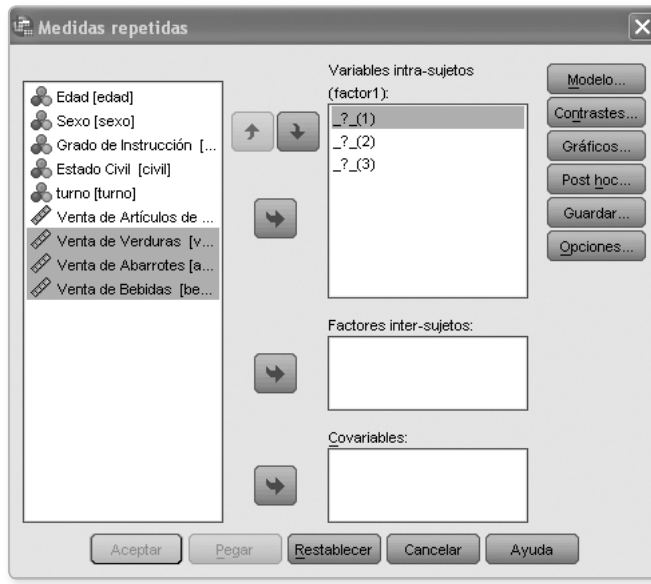
Cuando se selecciona el procedimiento *Medidas repetidas,* se muestra un cuadro de diálogo en el cual se indica cuántas variables se van a comparar. En este caso son tres variables, luego:




Define el número de niveles y pulse en Añadir (Add). A continuación, pulse en Definir.

A continuación, se pulsa sobre el botón Añadir para que el SPSS interprete que se ha definido un factor *intrasujetos* (denominado, en el ejemplo, Factor 1) con tres niveles diferentes. Después de Añadir, se pulsa el botón Definir para indicar al SPSS cuáles son las variables que forman cada uno de los niveles de Factor 1, en este caso tres niveles.

Al igual que en otros procedimientos, el SPSS muestra un cuadro de diálogo con todas las variables del archivo de datos. En este cuadro, se seleccionan las variables que forman los grupos del factor intrasujetos que, en el ejemplo, son *Venta de verduras, Venta de abarrotes* y *Venta de bebidas.*




Seguidamente, pulsando sobre el botón flecha y se trasladan estas variables al cuadro denominado «Variables intrasujetos». De este modo, se ha indicado cuáles son los tres niveles del factor intrasujetos **Factor 1**.

El SPSS proporciona distintas tablas de resultados. La tabla correspondiente a la F del factor intrasujetos se denomina «Pruebas de efectos intrasujetos»:

**Tests of Within-Subjects Effects**

Measure: MEASURE_1

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| FACTOR1 | Sphericity Assumed | 14768.339 | 2 | 7384.169 | 513.974 | .000 |
| | Greenhouse-Geisser | 14768.339 | 1.404 | 10517.592 | 513.974 | .000 |
| | Huynh-Feldt | 14768.339 | 1.454 | 10156.961 | 513.974 | .000 |
| | Lower-bound | 14768.339 | 1.000 | 14768.339 | 513.974 | .000 |
| Error(FACTOR1) | Sphericity Assumed | 833.275 | 58 | 14.367 | | |
| | Greenhouse-Geisser | 833.275 | 40.721 | 20.463 | | |
| | Huynh-Feldt | 833.275 | 42.166 | 19.762 | | |
| | Lower-bound | 833.275 | 29.000 | 28.734 | | |

La tabla contiene la suma de cuadrados intergrupos y error con los correspondientes grados de libertad, medias cuadráticas, el estadístico de contraste F y el nivel crítico. El SPSS ofrece los resultados bajo varios estadísticos, según se cumpla o no el supuesto de esfericidad.

Como puede verse en el cuadro de diálogo y en el visor de resultados, este procedimiento tiene muchas más opciones, además de las indicadas. Por ejemplo, pueden realizarse pruebas de tendencia, comparaciones entre medias, gráficos, etc. Asimismo, desde este menú también puede realizar ANOVA con medidas repetidas en más de un factor y con medidas repetidas solo en algunos de ellos (diseños mixtos).

### 5.2.3. EJERCICIOS PROPUESTOS

1. En el ejercicio propuesto número 3 de comparación de medias, supóngase que se vuelven a medir las horas trabajadas por los sujetos transcurrido un mes, y se obtienen los siguientes resultados:

| Sujeto | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Horas | 7.53 | 6.42 | 9.39 | 7.18 | 7.49 | 9.20 | 7.24 | 7.11 | 8.37 | 6.29 |

2. Según estos datos, ¿puede afirmarse que la terapia sigue teniendo efecto transcurrido un mes, con un nivel de confianza del 95%?

3. Un profesor de Estadística de una facultad evalúa su asignatura, a partir de tres controles que se realizan a lo largo del curso, y hacen media con la nota del examen final. El profesor desea saber si el rendimiento de los alumnos ha ido aumentando en cada uno de los controles. Para ello, selecciona aleatoriamente una muestra de cinco alumnos. Sus calificaciones obtenidas en los tres controles se muestran en la siguiente tabla:

| | | | | | |
|---|---|---|---|---|---|
| Control 1: | 7 | 4 | 5 | 3 | 2 |
| Control 2 | 5 | 6 | 6 | 4 | 3 |
| Control 3 | 5 | 8 | 6 | 5 | 4 |

   a. Indique las variables que intervienen en el problema y la función que desempeñan.
   b. Seleccione el modelo de ANOVA más apropiado, plantee la hipótesis pertinente y tome una decisión con  = 0,05.
   c. ¿Existen diferencias entre las calificaciones en los controles?

4. Suponga que se dispone de los datos sobre si los sujetos considerados son fumadores (1) o no lo son (0).

| Sujeto | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fuma | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |

   Contrasta el efecto de las variables Edad y Fumar sobre el número de horas dormidas con =0,05.

5. Un gabinete de psicología clínica pretende estudiar la eficacia de cuatro terapias (psicoanalítica, conductista, cognitivista y gestáltica) en el tratamiento de los trastornos del sueño. Para ello, asigne aleatoriamente a un grupo de 24 pacientes (mitad varones, mitad mujeres) a cada terapia y mida las horas que duermen, transcurrido un mes después de la terapia. Los resultados obtenidos son los siguientes:

| | Varones | | | Mujeres | | |
|---|---|---|---|---|---|---|
| Psicoanalítica | 6 | 4 | 3 | 3 | 4 | 3 |
| Conductista | 7 | 7 | 8 | 5 | 6 | 5 |
| Cognitivista | 9 | 8 | 7 | 6 | 7 | 7 |
| Gestáltica | 5 | 6 | 5 | 4 | 5 | 3 |

6. ¿Qué puede concluirse con =0,01?

7. ¿Qué terapia recomendaría a un paciente que acudiese a su consulta con insomnio?

8. Represente gráficamente la interacción, e interprete el resultado.

9. Con los datos del archivo supermercado, contraste el efecto de las variables Venta de abarrotes y Sexo sobre el gasto con =0,05.

## 5.3. CORRELACIÓN

El procedimiento Correlaciones permite calcular la matriz de correlaciones de un conjunto de variables. **Correlación** permite obtener los índices de asociación lineal con variables cuantitativas; asimismo, permite evaluar la existencia de relaciones lineales entre variables. Sin necesidad de calcular algún estadístico, es posible elaborar un gráfico de dispersión que informe sobre la posibilidad que se dé este tipo de relación. La correlación presenta los procedimientos **Bivariadas** (Bivariate), **Parciales** (Partial) y **Distancias** (Distances), que se definen a continuación.




### A. Bivariadas

El procedimiento Bivariadas calcula el coeficiente de correlación de Pearson, la rho de Spearman y la tau-b de Kendall con sus niveles de significación. Las correlaciones miden cómo están relacionados las variables o los órdenes de los rangos. Antes de calcular un coeficiente de correlación, se deben inspeccionar los datos para detectar valores atípicos (que pueden producir resultados equívocos) y evidencias de una relación lineal.

Las correlaciones son medidas de asociación lineal. Dos variables pueden estar perfectamente relacionadas, pero si la relación no es lineal, el coeficiente de correlación no será un estadístico adecuado para medir su asociación.

### B. Parciales

El procedimiento Parciales calcula los coeficientes de correlación parcial, los cuales describen la relación lineal existente entre dos variables mientras se controlan los efectos de una o más variables adicionales.

### C. Distancias

El procedimiento Distancias calcula una gran variedad de estadísticos, midiendo las similaridades o las disimilaridades (distancias), bien entre pares de variables o entre pares de casos. Estas medidas de similaridad o de distancia se pueden emplear a continuación con otros procedimientos, como el análisis factorial, el análisis de conglomerados o el escalamiento multidimensional, para ayudar a analizar conjuntos de datos complejos.

## 5.3.1. EJERCICIO PRÁCTICO 1

En el presente tema, se estudiará el análisis de datos con dos o más variables y la forma de proceder con SPSS. En el ejercicio práctico, se aprenderá a utilizar los procedimientos de Correlación. Los datos utilizados serán los recopilados en la encuesta del supermercado presentado en el capítulo dos. Aprendiendo a realizar análisis con Correlación. Como se mencionó anteriormente, todos los análisis se realizan desde el menú Analizar > Correlación (Correlate).



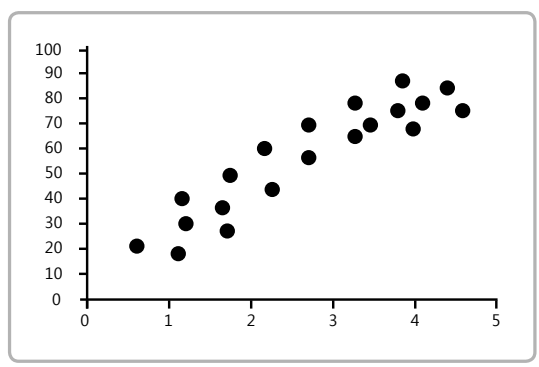

Para realizar análisis con los procedimientos de Correlación (*Bivariadas, Parciales y Distancias*), se deben seguir los cuadros de diálogo. A continuación, se presenta una explicación detallada del cuadro de diálogo correspondiente a los principales procedimientos que le tocará analizar en este tema, como es la herramienta **Correlación**.

### A. Coeficiente de correlación lineal

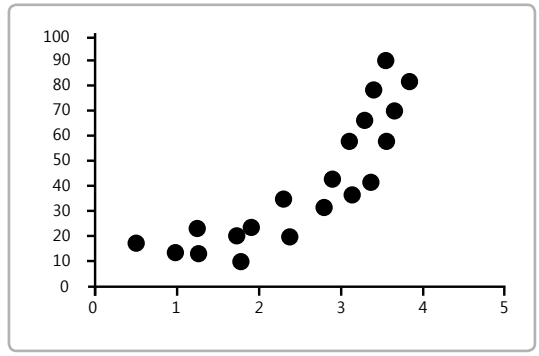
En una distribución bidimensional. Puede ocurrir que las dos variables guarden algún tipo de relación entre sí. Por ejemplo, si se analiza la estatura y el peso de los alumnos de una clase, es posible que exista relación entre ambas variables: mientras más alto sea el alumno, mayor será su peso.

El **coeficiente de correlación lineal** mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si se representara en un gráfico los pares de valores de las dos variables, la nube de puntos se aproximará a una recta).

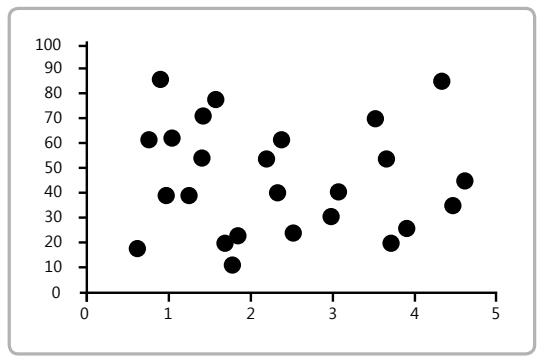
- Relación lineal



- Relación exponencial:




- Sin relación:




Cabe la posibilidad de que exista una relación que no sea lineal sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación entre las variables, por lo que convendría emplear otro tipo de coeficiente más apropiado. Por tanto, si se puede usar el coeficiente de correlación lineal, lo mejor es representar los pares de valores en un gráfico y observar qué forma describe.

Sea X: $x_1, x_2, \ldots x_n$ y Y: $y_1, y_2, \ldots y_n$, los valores de dos variables (X, Y). El coeficiente de correlación lineal se calcula aplicando la siguiente fórmula:

$$r_{xy} = \frac{Sxy}{\sqrt{S_x^2 * S_y^2}}$$

Donde:

$$Sxy = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{XY} \qquad \text{: Covarianza de (X,Y)}$$

$$S_x^2 = \frac{\sum_{i=1}^{n}\left[x_i - \overline{X}\right]^2}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 \qquad \text{: Varianza de X}$$

$$S_y^2 = \frac{\sum_{i=1}^{n}\left[y_i - \overline{Y}\right]^2}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2 \qquad \text{: Varianza de Y}$$

El numerador se denomina «covarianza», y el denominador calcula el producto de las varianzas de "x" y de "y"; a este producto se le calcula la raíz cuadrada.

**B. Propiedades del coeficiente de correlación lineal ($r_{xy}$)**

Los valores que puede tomar el coeficiente de correlación r están comprendidos en el intervalo: $-1 \leq r_{xy} \leq 1$.

**a. Si $r_{xy} > 0$, la correlación lineal es positiva**

Los valores positivos indican que las dos variables aumentan o disminuyen al mismo tiempo (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se aproxime a 1. Por ejemplo, altura y peso: los alumnos más altos suelen pesar más.

**b. Si $r_{xy} < 0$, la correlación lineal es negativa**

Valores negativos significa que, cuando una variable aumenta, la otra disminuye o viceversa (si sube el valor de una variable, disminuye la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1. Por ejemplo, peso y velocidad: los alumnos más gordos suelen correr menos.

**c. Si $r_{xy} = 0$, no existe correlación lineal entre las variables**

Aunque podría existir otro tipo de correlación: parabólica, exponencial, etc. De todos modos, pese a que el valor de r fuese próximo a 1 o -1, tampoco implica necesariamente que existe una relación de causa-efecto entre las dos variables, ya que este resultado podría haberse debido al puro azar.

## Ejemplo 1:

Calcule el coeficiente de correlación lineal de la siguiente serie de datos de altura y peso de los alumnos de una clase:

| Alumno | Estatura | Peso | Alumno | Estatura | Peso | Alumno | Estatura | Peso |
|---|---|---|---|---|---|---|---|---|
| Alumno 1 | 1,25 | 32 | Alumno 11 | 1,25 | 33 | Alumno 21 | 1,25 | 33 |
| Alumno 2 | 1,28 | 33 | Alumno 12 | 1,28 | 35 | Alumno 22 | 1,28 | 34 |
| Alumno 3 | 1,27 | 34 | Alumno 13 | 1,27 | 34 | Alumno 23 | 1,27 | 34 |
| Alumno 4 | 1,21 | 30 | Alumno 14 | 1,21 | 30 | Alumno 24 | 1,21 | 31 |
| Alumno 5 | 1,22 | 32 | Alumno 15 | 1,22 | 33 | Alumno 25 | 1,22 | 32 |
| Alumno 6 | 1,29 | 35 | Alumno 16 | 1,29 | 34 | Alumno 26 | 1,29 | 34 |
| Alumno 7 | 1,30 | 34 | Alumno 17 | 1,30 | 35 | Alumno 27 | 1,30 | 34 |
| Alumno 8 | 1,24 | 32 | Alumno 18 | 1,24 | 32 | Alumno 28 | 1,24 | 31 |
| Alumno 9 | 1,27 | 32 | Alumno 19 | 1,27 | 33 | Alumno 29 | 1,27 | 35 |
| Alumno10 | 1,29 | 35 | Alumno 20 | 1,29 | 33 | Alumno 30 | 1,29 | 34 |

## Solución:

Aplique la fórmula:

$$r_{xy} = \frac{Sxy}{\sqrt{S_x^2 * S_y^2}} = \frac{(0{,}826 / 30)}{\sqrt{\left(\frac{0{,}02568}{30}\right) * \left(\frac{51{,}366}{30}\right)}} = 0.8281$$

$r_{xy} = 0.8281$

Por lo tanto, la correlación existente entre estas dos variables es elevada (82.8%) y de signo positivo; es decir, a más estatura, más peso.

Resultados con SPSS:

**Correlaciones**

| | | Estatura | Peso |
|---|---|---|---|
| Estatura | Correlación de Pearson | 1 | ,828** |
| | Sig. (unilateral) | | ,000 |
| | N | 30 | 30 |
| Peso | Correlación de Pearson | ,828** | 1 |
| | Sig. (unilateral) | ,000 | |
| | N | 30 | 30 |

**. La correlación es significante al nivel 0,01 (unilateral).

Se puede observar que SPSS presenta las correlaciones en forma matricial, tal como debe ser. En el ejemplo, "x" es la estatura e "y" es el peso.

$$R = \begin{bmatrix} r_{xx} & r_{xy} \\ r_{xy} & r_{yy} \end{bmatrix} = \begin{bmatrix} 1 & 0.828 \\ 0.828 & 1 \end{bmatrix}$$

La Correlación de la estatura versus el peso es igual a 0.828 (82.8%); mientras que la correlación de la estatura versus la estatura es igual a 1; es decir, la correlación de una misma variable siempre será igual a 1.

## C. Coeficiente de correlación lineal con SPSS bivariadas

### a. Índice de asociación lineal

El índice de asociación lineal es medido por el procedimiento Bivariada. Para obtener los índices de asociación lineal con variables (la covarianza $S_{XY}$ y el coeficiente de correlación de Pearson $r_{XY}$) se utiliza el procedimiento bivariadas, se selecciona el menú **Analizar > Correlaciones > Bivariadas**.

El cuadro de diálogo para el procedimiento Correlaciones bivariadas tiene el siguiente aspecto:

Pulse para ver opciones.

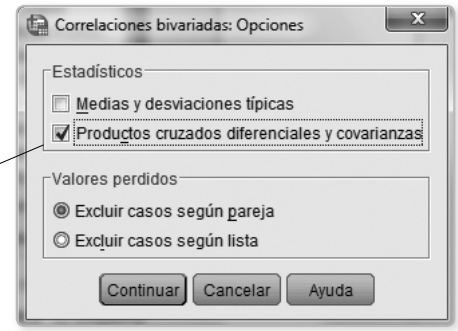
Traslade a este cuadro las variables para las que se desee obtener un coeficiente de correlación lineal.

Desde aquí, se pueden obtener el coeficiente de Pearson ($r_{XY}$), Kendall y Spearman. Los dos últimos sirven para variables ordinales.



Si pulsa el botón Opciones, se muestra el siguiente cuadro de diálogo:

Seleccione si desea obtener la matriz de varianzas-covarianzas.




Desde aquí, puede indicarse que se muestren los estadísticos descriptivos (media y desviación típica) para cada una de las variables seleccionadas, y también la matriz de varianzas-covarianzas (selección Productos cruzados diferenciales y covarianzas).

### b. Matriz de varianzas-covarianzas

Este procedimiento se consigue pulsando opciones, luego marque varianzas-covarianzas (*cross-product deviations and covariances*). En la covarianza ocurre como en la varianza: el estadístico ofrecido es el insesgado. En este caso, se tiene que:

$$S'_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1}$$

### c. Matriz de correlaciones

Desde el cuadro de diálogo anterior, una vez asignadas las variables seleccionadas al cuadro Variables, si se pulsa Aceptar, la matriz de correlaciones de Pearson aparecerá entre todas las posibles variables seleccionadas en el Visor de resultados.

Observe cómo se resuelve en SPSS alguno de los ejemplos. Al parecer, los datos de 30 sujetos para dos variables: *X* (Venta de verduras) e *Y* (Venta de abarrotes). Finalmente, se pulsa el botón Aceptar. Los resultados obtenidos son los siguientes:

**Correlations**

| | | Venta de Verduras | Venta de Abarrotes |
|---|---|---|---|
| Venta de Verduras | Pearson Correlation | 1.000 | -.187 |
| | Sig. (2-tailed) | . | .323 |
| | Sum of Squares and Cross-products | 949.467 | -71.340 |
| | Covariance | 32.740 | -2.460 |
| | N | 30 | 30 |
| Venta de Abarrotes | Pearson Correlation | -.187 | 1.000 |
| | Sig. (2-tailed) | .323 | . |
| | Sum of Squares and Cross-products | -71.340 | 153.372 |
| | Covariance | -2.460 | 5.289 |
| | N | 30 | 30 |

Se ha señalado el coeficiente de Pearson con un círculo. Como se observa, se obtiene $r_{XY}$ = -0.187. La media y la desviación estándar se muestran en la tabla siguiente:

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Venta de Verduras | 12.5333 | 5.7219 | 30 |
| Venta de Abarrotes | 34.5400 | 2.2997 | 30 |

Vea un ejemplo a partir de los datos de la encuesta. Tome las variables Venta de aseo personal, Venta de verduras y Venta de bebidas, obteniéndose sus coeficientes de correlación. Los resultados son los siguientes:

**Correlations**

| | | Venta de Artículos de Aseo Personal | Venta de Verduras | Venta de Bebidas |
|---|---|---|---|---|
| Venta de Artículos de Aseo Personal | Pearson Correlation | 1.000 | -.208 | -.154 |
| | Sig. (2-tailed) | . | .271 | .418 |
| | Sum of Squares and Cross-products | 146.595 | -77.453 | -17.667 |
| | Covariance | 5.055 | -2.671 | -.609 |
| | N | 30 | 30 | 30 |
| Venta de Verduras | Pearson Correlation | -.208 | 1.000 | .182 |
| | Sig. (2-tailed) | .271 | . | .335 |
| | Sum of Squares and Cross-products | -77.453 | 949.467 | 53.333 |
| | Covariance | -2.671 | 32.740 | 1.839 |
| | N | 30 | 30 | 30 |
| Venta de Bebidas | Pearson Correlation | -.154 | .182 | 1.000 |
| | Sig. (2-tailed) | .418 | .335 | . |
| | Sum of Squares and Cross-products | -17.667 | 53.333 | 90.167 |
| | Covariance | -.609 | 1.839 | 3.109 |
| | N | 30 | 30 | 30 |

**Descriptive Statistics**

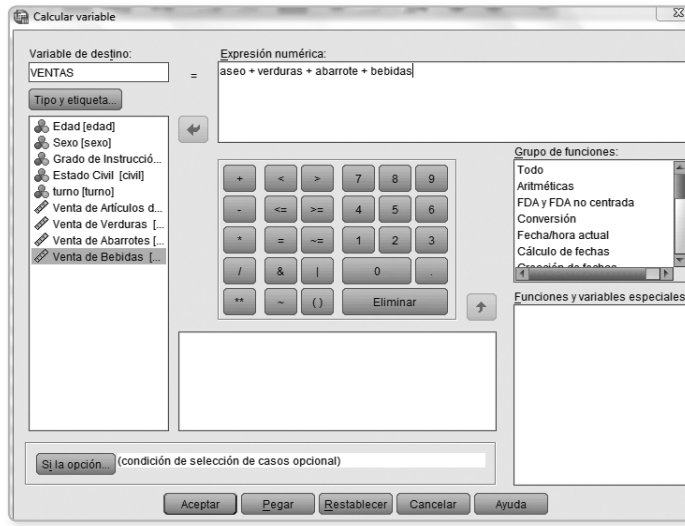
| | Mean | Std. Deviation | N |
|---|---|---|---|
| Venta de Artículos de Aseo Personal | 2.6533 | 2.2483 | 30 |
| Venta de Verduras | 12.5333 | 5.7219 | 30 |
| Venta de Bebidas | 4.1667 | 1.7633 | 30 |

Se han sombreado los tres coeficientes de Pearson. Como se observa, no se ha obtenido un valor razonable de correlación lineal; esto es, las correlaciones son muy bajas.

### D. Combinación lineal de variables

En este apartado se verá cómo se realizan transformaciones lineales con los datos de dos o más variables para obtener nuevas puntuaciones (por ejemplo: T*i* = X*i* + Y*i*; T*i* = AX*i* + BY*i*; etc.), y cómo se cumplen las propiedades sobre la media y la varianza en estas nuevas variables.

El primer paso es definir la nueva variable T*i*. Para ello, se emplea el menú Transformar > Calcular (ver § IV). El segundo paso es dar un nombre a la nueva variable y, a continuación, indicar la expresión numérica que la define. Con los datos de la encuesta del supermercado, cree una combinación lineal con las cuatro variables de ventas (Útiles de aseo personal, Verduras, Abarrotes y bebidas). A la variable resultante se le llamará «Ventas». El cuadro de diálogo de Transformar > Calcular queda del siguiente modo:




La nueva variable es del tipo *Ti* = *Xi* + *Yi* + *Vi* + *Wi*. Ahora se obtienen sus estadísticos descriptivos univariados desde el menú Analizar > Estadísticos descriptivos > Descriptivos y las covarianzas desde **Analizar > Correlaciones > Bivariadas (Opciones)**. El resultado es el siguiente:

**Descriptive Statistics**

| | N | Mean | Variance |
|---|---|---|---|
| Venta de Artículos de Aseo Personal | 30 | 2.6533 | 5.055 |
| Venta de Verduras | 30 | 12.5333 | 32.740 |
| Venta de Abarrotes | 30 | 34.5400 | 5.289 |
| Venta de Bebidas | 30 | 4.1667 | 3.109 |
| VENTAS | 30 | 53.8933 | 37.361 |
| Valid N (listwise) | 30 | | |

**Correlations**

| | | Venta de Artículos de Aseo Personal | Venta de Verduras | Venta de Abarrotes | Venta de Bebidas |
|---|---|---|---|---|---|
| Venta de Artículos de Aseo Personal | Pearson Correlation | 1.000 | -.208 | .160 | -.154 |
| | Sig. (2-tailed) | . | .271 | .399 | .418 |
| | Sum of Squares and Cross-products | 146.595 | -77.453 | 23.966 | -17.667 |
| | Covariance | 5.055 | -2.671 | .826 | -.609 |
| | N | 30 | 30 | 30 | 30 |
| Venta de Verduras | Pearson Correlation | -.208 | 1.000 | -.187 | .182 |
| | Sig. (2-tailed) | .271 | . | .323 | .335 |
| | Sum of Squares and Cross-products | -77.453 | 949.467 | -71.340 | 53.333 |
| | Covariance | -2.671 | 32.740 | -2.460 | 1.839 |
| | N | 30 | 30 | 30 | 30 |
| Venta de Abarrotes | Pearson Correlation | .160 | -.187 | 1.000 | -.331 |
| | Sig. (2-tailed) | .399 | .323 | . | .074 |
| | Sum of Squares and Cross-products | 23.966 | -71.340 | 153.372 | -38.900 |
| | Covariance | .826 | -2.460 | 5.289 | -1.341 |
| | N | 30 | 30 | 30 | 30 |
| Venta de Bebidas | Pearson Correlation | -.154 | .182 | -.331 | 1.000 |
| | Sig. (2-tailed) | .418 | .335 | .074 | . |
| | Sum of Squares and Cross-products | -17.667 | 53.333 | -38.900 | 90.167 |
| | Covariance | -.609 | 1.839 | -1.341 | 3.109 |
| | N | 30 | 30 | 30 | 30 |

Como se observa, la media en la nueva variable es 53.8933, valor que se corresponde exactamente con el resultado de la suma de las medias en las cuatro variables originales. En cuanto a la varianza, el resultado es 37.361.

### 5.3.2. EJERCICIO PRÁCTICO 2

Los datos para este ejercicio aparecen recopilados en las variables *Día1, Día2* y *Día3,* que indican el número de horas trabajadas en los tres días.

Además, se ha recopilado la variable Sexo que toma valores 0 para mujeres y 1 para hombres. *Edad*, 1 para adolescentes, 2 para jóvenes y 3 para adultos.

Finalmente, se incluyen tres datos relacionados con las terapias que ha recibido el sujeto. Todas ellas toman 1 si el sujeto ha recibido la terapia y 0 en caso contrario. Las variables son *Estrés, Ansiedad y Fobia*.

| Sujeto | Día1 | Día2 | Día3 | Sexo | Edad | Estrés | Ansiedad | Fobia |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.7 | 6.3 | 6.8 | 0 | 1 | 0 | 1 | 0 |
| 2 | 4.8 | 5.4 | 4.3 | 1 | 3 | 0 | 0 | 0 |
| 3 | 7.6 | 7.9 | 8.6 | 0 | 2 | 0 | 0 | 0 |
| 4 | 5.7 | 6.0 | 6.7 | 1 | 1 | 0 | 0 | 1 |
| 5 | 3.8 | 4.6 | 4.9 | 1 | 3 | 0 | 0 | 0 |
| 6 | 7.5 | 8.0 | 8.5 | 1 | 2 | 0 | 1 | 0 |
| 7 | 6.4 | 7.0 | 7.3 | 1 | 1 | 1 | 1 | 0 |
| 8 | 7.7 | 8.1 | 8.5 | 0 | 2 | 0 | 0 | 0 |
| 9 | 4.0 | 4.5 | 5.0 | 0 | 3 | 1 | 0 | 0 |
| 10 | 5.7 | 6.1 | 6.8 | 1 | 2 | 1 | 0 | 1 |

#### A. Correlaciones con SPSS

El procedimiento Correlaciones permite calcular la matriz de correlaciones de un conjunto de variables. Para calcular la correlación de Pearson, se ejecuta el procedimiento **Analizar > Correlaciones > Bivariadas**. Por ejemplo, para calcular las correlaciones entre los tres días, se indica lo siguiente:



Pulsando el botón Opciones pueden obtenerse los estadísticos descriptivos y la matriz de covarianzas. Según puede verse en el cuadro de diálogo, este procedimiento también permite calcular la correlación de tau de Kendall y la de Spearman, que utilizan únicamente las propiedades ordinales de los datos. Al pulsar en Aceptar, el SPSS muestra la matriz de correlaciones entre las tres variables:

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Día 1 | 5.8900 | 1.4255 | 10 |
| Día 2 | 6.3900 | 1.3404 | 10 |
| Día 3 | 6.7400 | 1.5756 | 10 |

**Correlations**

| | | Día 1 | Día 2 | Día 3 |
|---|---|---|---|---|
| Día 1 | Pearson Correlation | 1.000 | .996** | .953** |
| | Sig. (2-tailed) | . | .000 | .000 |
| | Sum of Squares and Cross-products | 18.289 | 17.119 | 19.264 |
| | Covariance | 2.032 | 1.902 | 2.140 |
| | N | 10 | 10 | 10 |
| Día 2 | Pearson Correlation | .996** | 1.000 | .946** |
| | Sig. (2-tailed) | .000 | . | .000 |
| | Sum of Squares and Cross-products | 17.119 | 16.169 | 17.984 |
| | Covariance | 1.902 | 1.797 | 1.998 |
| | N | 10 | 10 | 10 |
| Día 3 | Pearson Correlation | .953** | .946** | 1.000 |
| | Sig. (2-tailed) | .000 | .000 | . |
| | Sum of Squares and Cross-products | 19.264 | 17.984 | 22.344 |
| | Covariance | 2.140 | 1.998 | 2.483 |
| | N | 10 | 10 | 10 |

**. Correlation is significant at the 0.01 level (2-tailed).

Además de los coeficientes de correlación de Pearson ($r_{ij}$) para cada combinación de las variables incluidas, la tabla contiene el nivel crítico que permite tomar una decisión respecto a cada hipótesis nula:

$$H_0: \rho_{ij} = 0.000$$

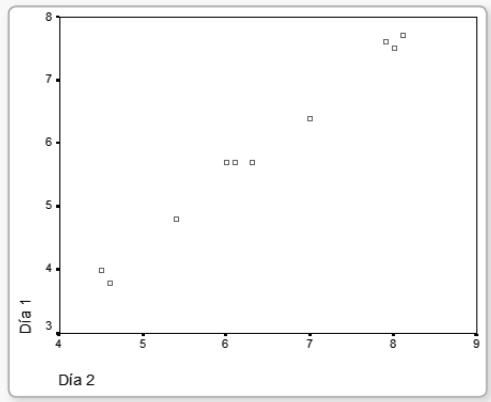
La última fila de la tabla contiene el número de casos (10).

### B. Correlación en gráficos

Las correlaciones lineales permiten evaluar la existencia de relaciones lineales entre variables. Sin necesidad de calcular algún estadístico, es posible elaborar un gráfico de dispersión que informe sobre la posibilidad que se dé este tipo de relación.

**Ejemplo 1:**

Use los datos del ejercicio práctico *2*. Mediante el procedimiento **Gráficos** > **Dispersión** > **Simples**, es posible elaborar el siguiente gráfico de dispersión de las variables: número de horas trabajadas durante los *Día 1* y *Día 2*.



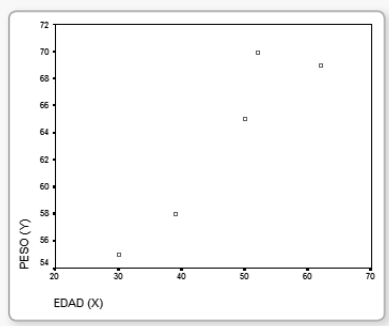

Cada punto del gráfico se corresponde con las puntuaciones de un sujeto en ambas variables. El gráfico sugiere que existe una relación lineal directa (o positiva) entre ambas, de modo que los sujetos que trabajan más el primer día también lo hacen el segundo. Para comprobar si esta relación es estadísticamente significativa, se utilizan los procedimientos **Correlaciones**.

## Ejercicio 2:

Se tienen los siguientes datos:

| Sujetos | Edad | Peso |
|---|---|---|
| 1 | 39 | 58 |
| 2 | 30 | 55 |
| 3 | 50 | 65 |
| 4 | 52 | 70 |
| 5 | 62 | 69 |

Los datos corresponden a cinco sujetos medidos en dos variables: Edad y Peso. Como en otros ejemplos previos, introduzca los datos correspondientes. Para realizar el diagrama de dispersión, seleccione el menú Gráficos de la barra de menús del Editor de datos, y elija el procedimiento Dispersión (Scatter). Luego, pulse Definir para definir los ejes del diagrama de las variables. Con esta definición, se obtiene el siguiente diagrama de dispersión simple en el Visor de resultados:




Como se observa en este diagrama, su forma indica que los puntos no están perfectamente alineados, pero se acercan a una hipotética línea recta.

CAPÍTULO

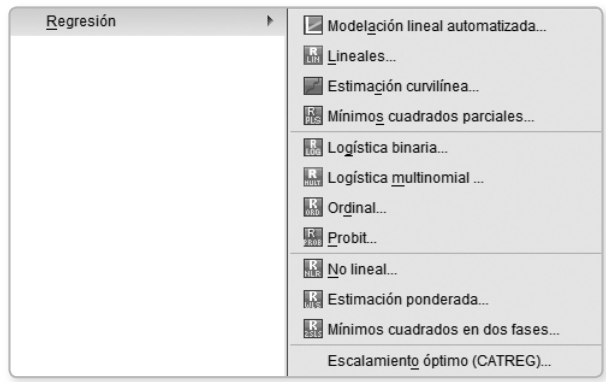
6

# REGRESIÓN LINEAL Y LOGLINEAL

Los temas tratados en el presente capítulo son: análisis de regresión lineal simple, diagrama de dispersión, ecuación de regresión, la ecuación de regresión con SPSS, bondad de ajuste del modelo, análisis de regresión lineal múltiple y loglineal.

## 6.1. REGRESIÓN

El **análisis de regresión** es una técnica estadística que se usa para estudiar la relación entre variables. Los principales procedimientos de la herramienta regresión son los siguientes:




### A. Lineal

Estima los coeficientes de la ecuación lineal, con una o más variables independientes que mejor predicen el valor de la variable dependiente.

### B. Estimación curvilínea

Genera estadísticos de estimación curvilínea por regresión, y gráficos relacionados para once modelos diferentes. Se produce un modelo diferente para cada variable dependiente; además, se pueden guardar valores pronosticados, residuos e intervalos pronosticados como nuevas variables.

### C. Logística binaria

Es útil cuando se desea predecir la presencia o ausencia de una característica o resultado, según los valores de un conjunto de variables predictoras. Es similar a un modelo de regresión lineal, pero está adaptada a modelos en los que la variable dependiente es dicotómica.

Los coeficientes pueden utilizarse para estimar la razón de las ventajas (*odds ratio*) de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante. Se pueden estimar modelos usando la entrada en bloque de las variables o cualquiera de los siguientes métodos por pasos.

### D. Condicional hacia adelante

- LR hacia adelante
- Wald hacia adelante
- Condicional hacia atrás
- LR hacia atrás
- Wald hacia atrás

### E. Logística multinomial

Resulta útil si se desea clasificar a los sujetos según los valores de un conjunto de variables predictoras. Este tipo de regresión es similar a la regresión logística, pero más general, ya que la variable dependiente no está restringida a dos categorías. Se ajusta un modelo logit multinomial para el modelo factorial completo o para un modelo especificado por el usuario; así, la estimación de los parámetros se realiza a través de un algoritmo iterativo de máxima verosimilitud.

### F. Ordinal

Permite dar forma a la dependencia de una respuesta ordinal politómica sobre un conjunto de predictores, que pueden ser factores o covariables. El diseño de la regresión ordinal se basa en la metodología de McCullagh, y en la sintaxis se hace referencia al procedimiento como PLUM.

### G. Probit

Mide la relación entre la intensidad de un estímulo y la proporción de casos que presentan una cierta respuesta a dicho estímulo. Es útil para las situaciones en las que se dispone de una respuesta dicotómica, que se piensa puede estar influenciada o causada por los niveles de alguna o algunas variables independientes, y es particularmente adecuada para datos experimentales. Este procedimiento permite estimar la intensidad necesaria para que un estímulo llegue a inducir una determinada proporción de respuestas, como la dosis efectiva para la mediana.

### H. No lineal

Sirve para encontrar un modelo no lineal para la relación entre la variable dependiente y un conjunto de variables independientes. A diferencia de la regresión lineal tradicional, que está restringida a la estimación de modelos lineales, la regresión no lineal puede estimar modelos con relaciones arbitrarias entre las variables independientes y la dependiente. Esto se lleva a cabo usando algoritmos de estimación iterativos. Este procedimiento no es necesario para modelos polinómicos simples, que se pueden estimar usando métodos tradicionales como la *regresión lineal*.

### I. Estimación ponderada

Los modelos de regresión lineal típicos asumen que la varianza es constante en la población objeto de estudio. Cuando este no es el caso (por ejemplo, cuando los casos con puntuaciones mayores en un atributo muestran más variabilidad que los casos con puntuaciones menores en ese atributo), la regresión lineal mediante mínimos cuadrados ordinarios (MCO, OLS) deja de proporcionar estimaciones óptimas para el modelo. Si las diferencias de variabilidad se pueden pronosticar a partir de otra variable, el procedimiento *Estimación ponderada* permite calcular los coeficientes de un modelo de regresión lineal mediante mínimos cuadrados ponderados (MCP, WLS), de forma que se les dé mayor ponderación a las observaciones más precisas (es decir, aquellas con menos variabilidad) al determinar los coeficientes de regresión. El procedimiento Estimación ponderada contrasta un rango de transformaciones de ponderación e indica cuál se ajustará mejor a los datos.

### J. Mínimos cuadrados en dos fases

Los modelos de regresión lineal típica asumen que los errores de la variable dependiente no están correlacionados con la variable o variables independientes. Cuando este no es el caso (por ejemplo, cuando las relaciones entre las variables son bidireccionales), la regresión lineal mediante mínimos cuadrados ordinarios (OLS) deja de proporcionar estimaciones óptimas del modelo.

La regresión por mínimos cuadrados en dos fases emplea variables instrumentales que no están correlacionadas con los términos de error para calcular los valores estimados de los predictores problemáticos (en la primera fase), y después usa dichos valores calculados para estimar un modelo de regresión lineal de la variable dependiente (la segunda fase). Dado que los valores calculados se basan en variables que no están correlacionadas con los errores, los resultados del modelo en dos fases son óptimos.

### K. Escalamiento óptimo

Cuantifica los datos categóricos mediante la asignación de valores numéricos a las categorías, obteniéndose una ecuación de regresión lineal óptima para las variables transformadas. El análisis de regresión lineal ordinario implica minimizar las diferencias de la suma de los cuadrados entre una variable de respuesta (la dependiente) y una combinación ponderada de las variables predictoras (las independientes). Las variables son normalmente cuantitativas, con los datos categóricos (nominales) recodificados como variables binarias o de contraste. Como resultado, las variables categóricas sirven para separar grupos de casos, y la técnica estima conjuntos separados de parámetros para cada grupo.

Los coeficientes estimados reflejan cómo los cambios en los predictores afectan a la respuesta. El pronóstico de la respuesta es posible para cualquier combinación de los valores predictores. Una aproximación alternativa incluye la regresión de la respuesta respecto a los propios valores predictores categóricos. Como consecuencia, se estima un coeficiente para cada variable; sin embargo, para las variables categóricas, los valores categóricos son arbitrarios. La codificación de las categorías de diferentes maneras proporciona diferentes coeficientes, dificultando las comparaciones entre los análisis de las mismas variables.

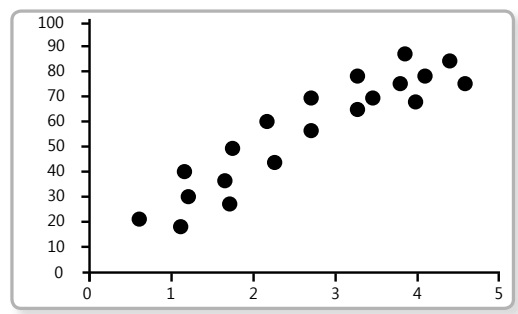
CATREG (regresión categórica mediante escalamiento óptimo) amplía la aproximación típica mediante un escalamiento de las variables nominales, ordinales y numéricas simultáneamente. El procedimiento cuantifica las variables categóricas, de manera que las cuantificaciones reflejen las características de las categorías originales. Además, trata a las variables categóricas cuantificadas como si fuesen variables numéricas. La utilización de transformaciones no lineales permite a las variables ser analizadas en varios niveles para encontrar el modelo que más se ajuste.

## 6.1.1. EJERCICIO PRÁCTICO 1

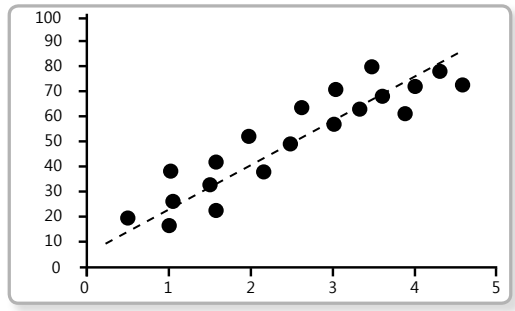
El ejercicio práctico consiste en realizar análisis de regresión por medio de los principales procedimientos de dicha herramienta, con los datos de la encuesta del supermercado.

### A. Regresión lineal simple

Represente en un gráfico los pares de valores de una distribución bidimensional: la variable X en el eje horizontal (eje de abscisa) y la variable Y en el eje vertical (eje de ordenada). Observe que la nube de puntos sigue una tendencia lineal:



El **coeficiente de correlación lineal** le permite determinar si efectivamente existe relación entre las dos variables. Una vez que se concluye que sí existe relación, la regresión le permite definir la recta que mejor se ajusta a esta nube de puntos.




El modelo de regresión lineal considera que la relación entre la variable dependiente Y (Y: $y_1$, $y_2$... $y_n$) y la independiente X (X: $x_1$, $x_2$... $x_n$) se puede formular a partir de la siguiente expresión lineal:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{... (1)}$$

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

A $\hat{y}_i$, se le llama «valor estimado»; luego $\hat{y}_i$ remplazando en la ecuación (1) se tiene que:

$$e_i = y_i - \hat{y}_i$$

Sea $\boldsymbol{e_i}$ la distancia (error o residuo) entre el valor observado $y_i$ su valor estimado $\hat{y}_i$. Para calcular los valores de los parámetros $\beta_0$ y $\beta_1$, use el método de mínimos cuadrados. El parámetro $\beta_1$ viene determinado por la siguiente fórmula:

$$\beta_1 = \frac{\sum_{i=1}^{n} \frac{(x_i - \overline{X})(y_i - \overline{Y})}{n}}{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n}} = \frac{S_{xy}}{S_x^2}$$

Es la covarianza de las dos variables x e y, dividida por la varianza de la variable x. El parámetro $\beta_0$ viene determinado por:

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

Es la media de la variable Y, menos la media de la variable X multiplicada por el parámetro $\boldsymbol{\beta_1}$ que se ha calculado.

El **parámetro $\boldsymbol{\beta_0}$** es el valor que toma la variable dependiente Y, cuando la variable independiente X vale y es el punto donde la recta cruza el eje vertical. El **parámetro $\boldsymbol{\beta_1}$** determina la pendiente de la recta y su grado de inclinación.

La **regresión lineal** le permite calcular el valor de estos dos parámetros, definiendo la recta que mejor se ajusta a esta nube de puntos.

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

**Ejemplo:**

Calcule la recta de regresión de la siguiente serie de datos de altura y peso de los alumnos de una clase. Considere que la altura es la variable independiente X, y que el peso es la variable dependiente Y (podría hacerlo también al contrario):

| Alumno | Estatura | Peso | Alumno | Estatura | Peso | Alumno | Estatura | Peso |
|---|---|---|---|---|---|---|---|---|
| Alumno 1 | 1,25 | 32 | Alumno 11 | 1,25 | 33 | Alumno 21 | 1,25 | 33 |
| Alumno 2 | 1,28 | 33 | Alumno 12 | 1,28 | 35 | Alumno 22 | 1,28 | 34 |
| Alumno 3 | 1,27 | 34 | Alumno 13 | 1,27 | 34 | Alumno 23 | 1,27 | 34 |
| Alumno 4 | 1,21 | 30 | Alumno 14 | 1,21 | 30 | Alumno 24 | 1,21 | 31 |
| Alumno 5 | 1,22 | 32 | Alumno 15 | 1,22 | 33 | Alumno 25 | 1,22 | 32 |
| Alumno 6 | 1,29 | 35 | Alumno 16 | 1,29 | 34 | Alumno 26 | 1,29 | 34 |
| Alumno 7 | 1,30 | 34 | Alumno 17 | 1,30 | 35 | Alumno 27 | 1,30 | 34 |
| Alumno 8 | 1,24 | 32 | Alumno 18 | 1,24 | 32 | Alumno 28 | 1,24 | 31 |
| Alumno 9 | 1,27 | 32 | Alumno 19 | 1,27 | 33 | Alumno 29 | 1,27 | 35 |
| Alumno10 | 1,29 | 35 | Alumno 20 | 1,29 | 33 | Alumno 30 | 1,29 | 34 |

El parámetro $\beta_1$ viene determinado por:

$$\beta_1 = (1 / 30) * 1{,}034 / ((1 / 30) * 0{,}00856) = 40{,}265$$

En tanto que, el parámetro $\beta_0$ por:

$$a = 33{,}1 - (40{,}265 * 1{,}262) = -17{,}714$$

Por lo tanto, la **recta** que mejor se ajusta a esta serie de datos es la siguiente:

$$Y = -17{,}714 + 40{,}265 * X$$

Resultados con SPSS:

**Resumen del modelo**

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación |
|---|---|---|---|---|
| 1 | ,828[a] | ,686 | ,675 | ,82519 |

a. Variables predictoras: (Constante), Estatura

**Coeficientes[a]**

| Modelo | | Coeficientes no estandarizados | | Coeficientes tipificados | | |
|---|---|---|---|---|---|---|
| | | B | Error típ. | Beta | t | Sig. |
| 1 | (Constante) | -17,714 | 6,500 | | -2,725 | ,011 |
| | Estatura | 40,265 | 5,149 | ,828 | 7,819 | ,000 |

**ANOVA[b]**

| Modelo | | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regresión | 41,634 | 1 | 41,634 | 61,142 | ,000[a] |
| | Residual | 19,066 | 28 | ,681 | | |
| | Total | 60,700 | 29 | | | |

a. Variables predictoras: (Constante), Estatura

b. Variable dependiente: Peso

### B. Análisis de regresión lineal simple con SPSS

El **análisis de regresión lineal** es una técnica estadística que se utiliza para estudiar la relación entre variables.

En la investigación estadística, suele emplearse para pronosticar valores de una variable criterio (*Y*), desde las puntuaciones de una variable predictora (*X*).

En el ejercicio se estudia la regresión lineal con una sola variable predictora (regresión simple); sin embargo, la regresión también puede efectuarse a partir de dos o más variables predictoras (*X*1, *X*2, etc.). A medida que se vaya ingresando en las opciones del SPSS para la regresión simple, se verá de forma intuitiva cómo se procede para la regresión con más de un predictor (regresión múltiple).

**NOTA**

*Antes de realizar un análisis de regresión, vea el comportamiento de los datos; la mejor manera es realizando un gráfico de dispersión.*

El ***diagrama de dispersión*** es una herramienta cuyo uso es necesario para poder explicar con claridad este modelo. La primera aproximación a la relación entre dos variables (*X* e *Y*) puede hacerse a partir de un diagrama de dispersión.

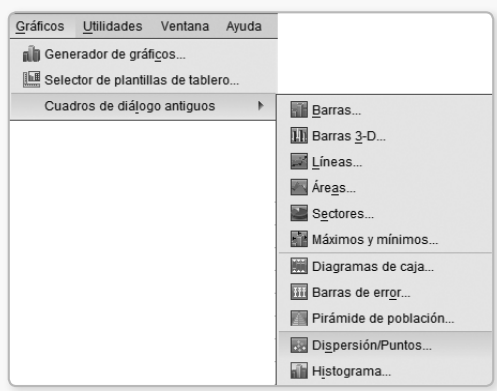
**Ejemplo:**

Se tienen los siguientes datos:

| Sujetos | Edad | Peso |
|---|---|---|
| 1 | 39 | 58 |
| 2 | 30 | 55 |
| 3 | 50 | 65 |
| 4 | 52 | 70 |
| 5 | 62 | 69 |

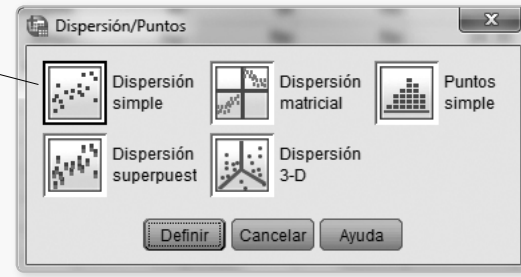
Los datos corresponden a cinco sujetos medidos en dos variables: Edad y Peso. Como en otros ejemplos previos, introduzca los datos correspondientes a los cinco sujetos en las dos variables en un archivo nuevo del Editor de datos. Una vez introducidos los datos, guárdelo como *regresion.sav*.

Para realizar el diagrama de dispersión se selecciona el menú Gráficos de la barra de menús del Editor de datos y se elige el procedimiento Dispersión (Scatter).



El cuadro de diálogo correspondiente es el siguiente:

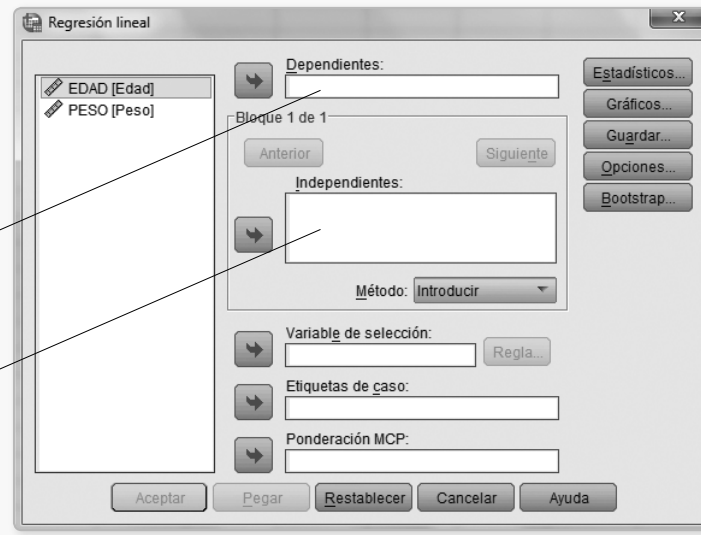
Pulse para definir los ejes del diagrama.



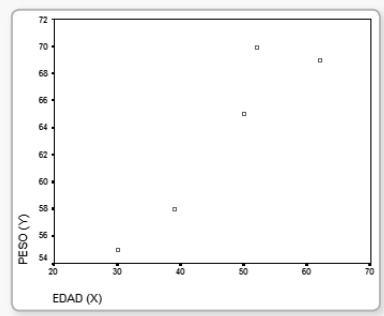

Luego, pulse Definir para definir los ejes del diagrama de las variables, obteniendo el cuadro de diálogo siguiente:

Variable Criterio: eje de ordenadas (Y)

Variable Predictora: eje de abscisas (X)




Con esta definición, se obtiene el siguiente diagrama de dispersión simple en el Visor de resultados:




Como se observa en este diagrama, su forma indica que los puntos en el diagrama no están perfectamente alineados, pero se acercan a una hipotética línea recta.

### C. Ecuación de regresión

Podrían trazarse diferentes rectas para realizar pronósticos de una variable, a partir de otra (por ejemplo, de *Y* a partir de *X* en una regresión de *Y* sobre *X*). Las rectas de regresión tienen una fórmula muy simple:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

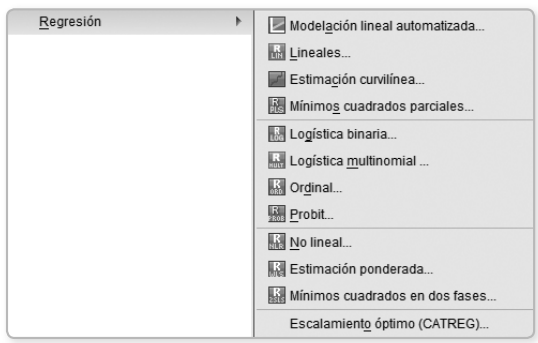
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

El objetivo es encontrar aquella recta que minimice la distancia entre lo encontrado (y) y lo pronosticado ($\hat{y}$); en otras palabras, que minimice la expresión $\sum(y_i - \hat{y}_i)/n$. Para ello, calcule los coeficientes del modelo mediante:

$$\beta_1 = \frac{\sum_{i=1}^{n} \frac{(x_i - \overline{X})(y_i - \overline{Y})}{n}}{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n}} = \frac{S_{xy}}{S_x^2}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

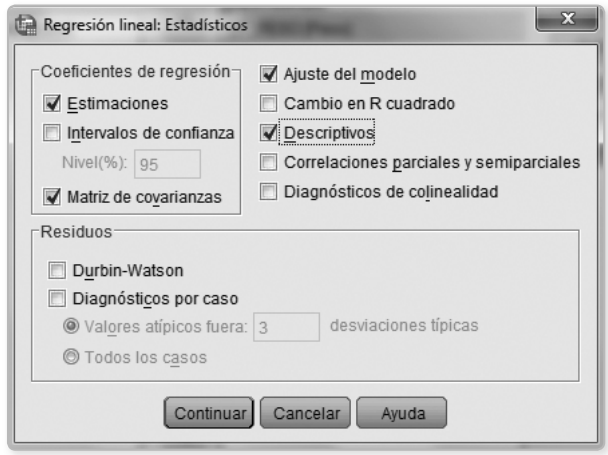
Para calcular dichos coeficientes en SPSS, debe obtener la recta de regresión y valorar la bondad del modelo; luego, se selecciona el menú Analizar > Regresión > Lineal.




El cuadro de diálogo correspondiente al procedimiento Regresión lineal aparece en la figura siguiente:



Como en otros cuadros de diálogo del SPSS, seleccione las variables. En este caso, debe distinguir entre *Dependiente* e *Independiente* (o independientes si se trata de una regresión múltiple). Dentro de este menú, hay otras opciones. Por el momento, se detallarán las opciones del botón Estadísticos que presenta el siguiente cuadro de diálogo:

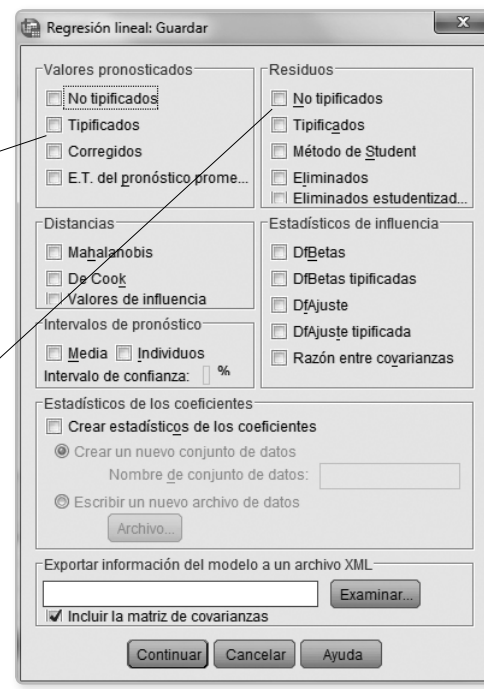



Desde aquí se solicita que se ofrezcan las estimaciones de la pendiente y el origen de la recta de regresión (según el criterio de mínimos cuadrados), la matriz de covarianzas para las variables, el ajuste del modelo (coeficiente de determinación) y los estadísticos descriptivos (media y varianza).

Otra opción relevante del menú regresión lineal es Guardar. Desde su correspondiente cuadro de diálogo, es posible indicar que se guarden los valores pronosticados por el modelo (las $y_i$) y los residuos (las $y_i - \hat{y}_i$) en el Editor de datos:

Seleccione si se desea guardar los valores pronosticados por el modelo (las $\hat{y}_i$) para cada uno de los sujetos.

Seleccione si se desea guardar lo no explicado por el modelo (las yi - $\hat{y}_i$) "los residuos" para cada uno de los sujetos.



Observe el resultado para el ejemplo dado. Si indica todas estas instrucciones, el resultado que ofrece el SPSS es el siguiente (nótese que se ofrecen muchas tablas de resultados; aquí se explican solo las más importantes).

**a. Regresión**

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| PESO (Y) | 63.4000 | 6.6558 | 5 |
| EDAD (X) | 46.6000 | 12.3612 | 5 |

**Correlations**

| | | PESO (Y) | EDAD (X) |
|---|---|---|---|
| Pearson Correlation | PESO (Y) | 1.000 | .935 |
| | EDAD (X) | .935 | 1.000 |
| Sig. (1-tailed) | PESO (Y) | . | .010 |
| | EDAD (X) | .010 | . |
| N | PESO (Y) | 5 | 5 |
| | EDAD (X) | 5 | 5 |

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | EDAD (X)[a] | . | Enter |

a. All requested variables entered.

b. Dependent Variable: PESO (Y)

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .935[a] | .875 | .833 | 2.7198 |

a. Predictors: (Constant), EDAD (X)

Esto es el coeficiente de determinación $R^2$ o la proporción de varianza en común entre X e Y.

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 155.008 | 1 | 155.008 | 20.954 | .020[a] |
| | Residual | 22.192 | 3 | 7.397 | | |
| | Total | 177.200 | 4 | | | |

a. Predictors: (Constant), EDAD (X)

b. Dependent Variable: PESO (Y)

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 39.932 | 5.269 | | 7.579 | .005 |
| | EDAD (X) | .504 | .110 | .935 | 4.578 | .020 |

a. Dependent Variable: PESO (Y)

En la columna B, se observan los coeficientes $\beta_0$ y $\beta_1$ del modelo, siendo $\beta_0$ = 39,932 y $\beta_1$ = 0,504. Luego, la recta estimada estará dada por:

$$Y = 39.932 + 0{,}504\ X$$

En la columna Beta, se muestra el modelo en típicas:

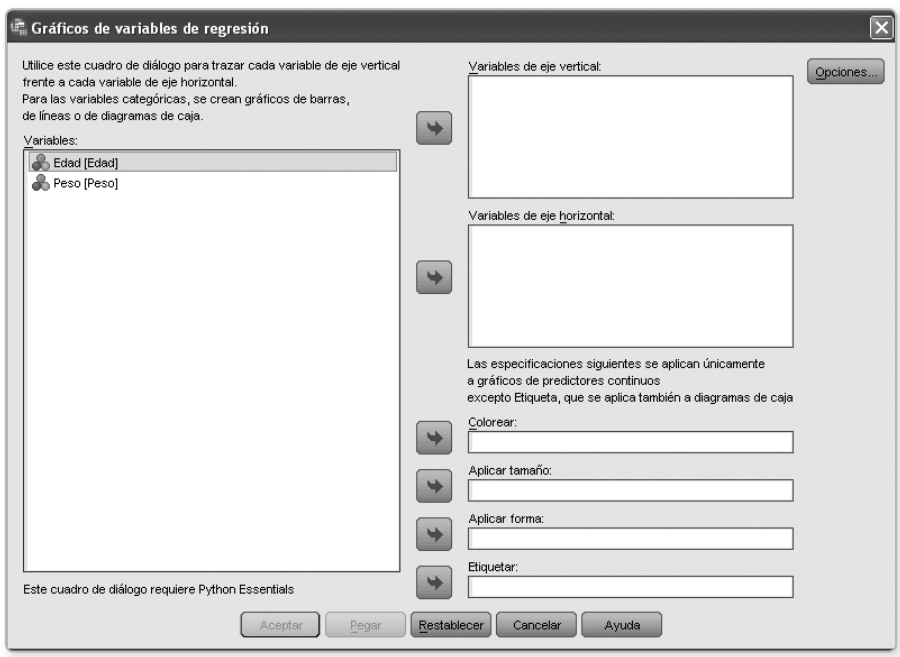
$$Z_{Y'} = r_{XY}\ Z_X \text{ (en el ejemplo: } Z_{Y'} = 0.504\ Z_X)$$

Como se observa, SPSS ofrece distintas tablas de resultados de la regresión; por el momento, solo nos fijaremos en dos: la primera, que se refiere a los coeficientes del modelo, y la segunda, referida a su bondad de ajuste. En cuanto a la primera, se toman los coeficientes no estandarizados. En este caso el mejor modelo para pronosticar *Yi* a partir de *Xi* es *Y'i* = 39.932 + 0,504 *Xi*.

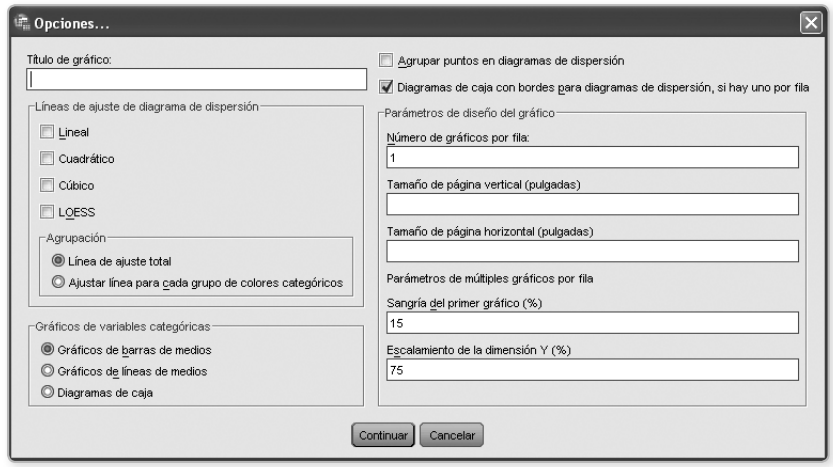
### D. Bondad de ajuste del modelo

La bondad de ajuste del modelo se mide mediante el coeficiente de determinación; sin embargo, resulta necesario disponer de información sobre el grado en que el modelo se ajusta a los datos, observados mediante una nube de puntos. En consecuencia, una primera aproximación es la interpretación gráfica del problema.

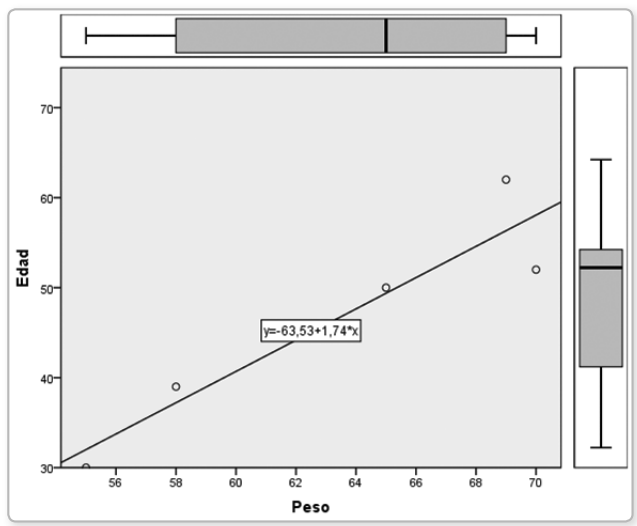
Para ilustrar esta técnica, se utilizan los datos *regresión.sav*. En primer lugar, se elaborará la gráfica del ajuste de la recta a los datos observados; se pulsa el menú **Gráficos** > **Gráficos de variable de regresión** y se abre el cuadro de diálogo de diagrama de dispersión interactiva.



En este cuadro, defina la variable del criterio (en el eje de ordenadas) y la variable predictora (en el eje de abscisas), desde la solapa Asignar variables. A continuación, seleccione Opciones, donde el cuadro de diálogo tiene el siguiente aspecto:




El gráfico obtenido es el siguiente:




Como se observa, los puntos no se alejan mucho de la recta, por lo que el ajuste es aceptable. Por otro lado, además de la interpretación gráfica del problema, la forma de cuantificar la bondad del modelo es mediante el coeficiente de determinación $r^2_{XY}$. Se trata de una medida estandarizada que toma valores entre 0 y 1, cuya interpretación es sencilla: representa la proporción de varianza explicada de la variable del criterio, a partir de la predictora.

**Resumen del modelo**

| Modelo | R | R cuadrado | R cuadrado ajustado | Error estándar de la estimación |
|---|---|---|---|---|
| 1 | ,935[a] | ,875 | ,833 | 5,051 |

a. Predictores: (Constante), Peso

En el caso del ejemplo, $r^2$ = 0,875, por lo que el *modelo sí es adecuado para hacer pronósticos de Y a partir de X*. La bondad del modelo también puede valorarse desde el análisis de los errores en los pronósticos, frecuentemente llamados «residuos» (Yi - Y'i). Nótese que, ejecutando las órdenes indicadas en el anterior apartado, han resultado dos nuevas variables en el Editor de datos (la variable 'pre_1' que se corresponde con las Y'i y lleva la etiqueta 'Unstandardized predicted value', y la variable 'res_1' que se refiere a las Yi - Y'i y lleva la etiqueta 'Unstandardized residual').

Teniendo los datos para Yi, Y'i e (Yi-Y'i) se puede evaluar la bondad del modelo, a partir de la descomposición de la varianza del criterio ($S^2_Y = S^2_{Y'} + S^2_{Y-Y'}$). Para ello, ingrese al menú Analizar > Estadísticos descriptivos > Descriptivos: Opciones e indique en el cuadro de diálogo que se calcule la varianza para estas tres variables (Y, pre_1 y res_1). Los resultados obtenidos en el visor son los siguientes:

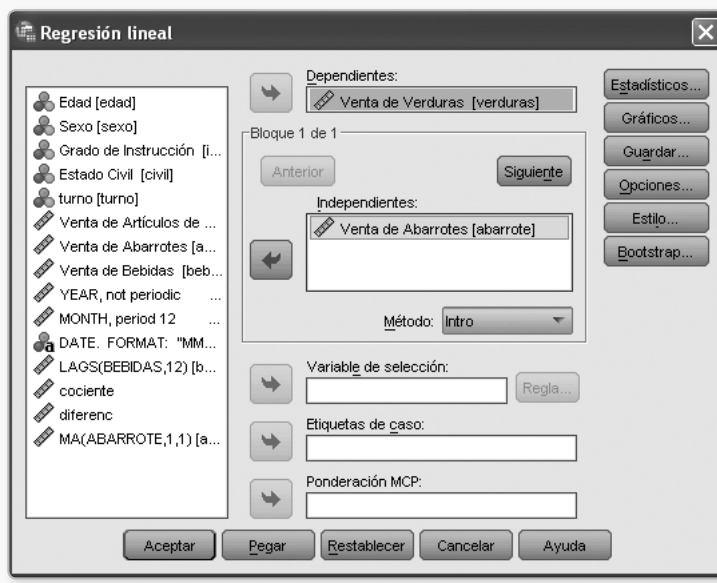
**Descriptive Statistics**

| | N | Variance |
|---|---|---|
| PESO (Y) | 5 | 44.300 |
| Unstandardized Predicted Value | 5 | 38.752 |
| Unstandardized Residual | 5 | 5.548 |
| Valid N (listwise) | 5 | |

Como se comprueba, la varianza del criterio ($S^2_Y$ = 44.3) se descompone en la varianza de los pronósticos ($S^2_{Y'}$ = 38.752) y la de los errores ($S^2_{Y-Y'}$= 5.548).

## Ejemplo 1:

A partir de los datos de la encuesta, se quiere predecir la variable *Venta de verduras* (V), a partir de la variable *Venta de abarrotes* (A). Para ello, construya la recta de regresión $V_i = \beta_0 + \beta_1 A_i$. La definición del modelo en el SPSS es la siguiente:




Los resultados obtenidos para el modelo planteado son los siguientes:

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Venta de Abarrotes[a] | . | Enter |

a. All requested variables entered.

b. Dependent Variable: Venta de Verduras

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .187[a] | .035 | .000 | 5.7205 |

a. Predictors: (Constant), Venta de Abarrotes

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 33.183 | 1 | 33.183 | 1.014 | .323[a] |
| | Residual | 916.283 | 28 | 32.724 | | |
| | Total | 949.467 | 29 | | | |

a. Predictors: (Constant), Venta de Abarrotes

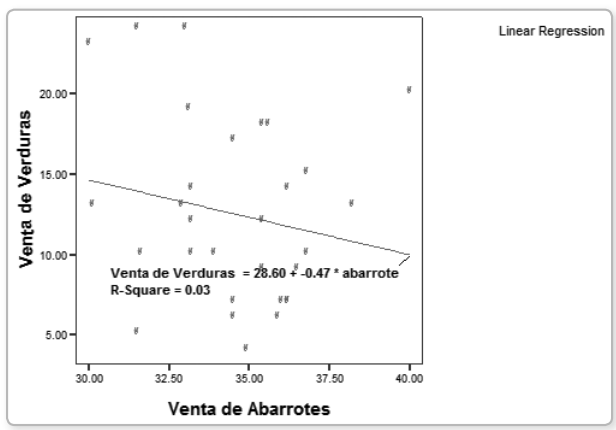
b. Dependent Variable: Venta de Verduras

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 28.599 | 15.989 | | 1.789 | .084 |
| | Venta de Abarrotes | -.465 | .462 | -.187 | -1.007 | .323 |

a. Dependent Variable: Venta de Verduras

El modelo resultante es Vi = 28.599 – 0.465Ai. Como se observa, el coeficiente de determinación (R al cuadrado) es 0,035, por lo que el modelo no es adecuado para explicar la relación entre la variable *Venta de verduras* y la variable *Venta de abarrotes*; es decir, esta última variable no tiene una capacidad predictiva para explicar la variable del criterio (Venta de verduras). Por otro lado, el gráfico obtenido define las opciones del diagrama de dispersión:



### E. Análisis de regresión lineal múltiple

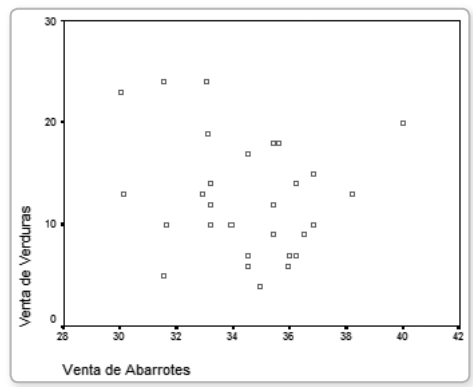
Una alternativa a esta situación de poco ajuste es incluir una variable predictora adicional (o más de una), y observar si su inclusión produce algún cambio en R cuadrado. Si considera dos variables predictoras, se está planteando el siguiente modelo:

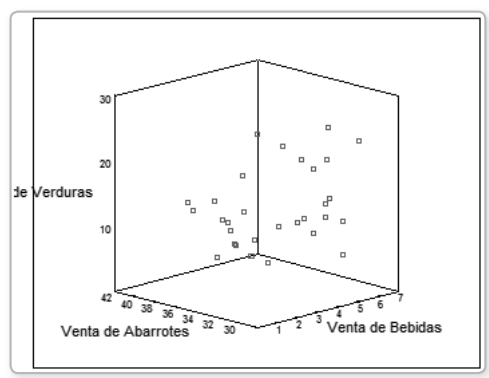
$$Yi = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

Retomando el ejemplo de la Venta de abarrotes y Venta de verduras, considerando ahora la Venta de bebidas (B) como segunda variable predictora. El modelo planteado es el siguiente:

$$V_i = \beta_0 + \beta_1 A_i . + \beta_2 B_i$$

La primera aproximación al problema puede hacerse a partir del gráfico de dispersión. En este caso, puede elaborarse el gráfico simple para observar la relación entre la edad y el peso, así como un gráfico 3D para observar la relación entre las dos predictoras y el criterio. El resultado obtenido estaría dado por:




Para obtener los coeficientes del modelo y valorar el ajuste del modelo, se selecciona el menú Analizar > Regresión > Lineal. Traslade las variables Venta de abarrotes y la Venta de bebidas a la casilla Independientes y la variable Venta de verduras a la casilla Dependiente. Los resultados encontrados son similares a los de la regresión simple, aunque ahora hay un coeficiente más que interpretar.

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Venta de Bebidas , Venta de Abarrotes[a] | . | Enter |

a. All requested variables entered.

b. Dependent Variable: Venta de Verduras

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .226[a] | .051 | -.019 | 5.7761 |

a. Predictors: (Constant), Venta de Bebidas , Venta de Abarrotes

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 48.648 | 2 | 24.324 | .729 | .492[a] |
| | Residual | 900.819 | 27 | 33.364 | | |
| | Total | 949.467 | 29 | | | |

a. Predictors: (Constant), Venta de Bebidas , Venta de Abarrotes

b. Dependent Variable: Venta de Verduras

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 22.926 | 18.168 | | 1.262 | .218 |
| | Venta de Abarrotes | -.354 | .494 | -.142 | -.716 | .480 |
| | Venta de Bebidas | .439 | .645 | .135 | .681 | .502 |

a. Dependent Variable: Venta de Verduras

Como se observa, el modelo resultante $V_i = 22.926 - 0.354\ A_i + 0.439\ B_i$ no aporta mucho a la explicación de la variable del criterio, debido a que el coeficiente de determinación ($R^2$) es 0.051 (5.1%).

En resumen, el modelo simple explicaba un 3,5 % ($R^2$=0.035) de varianza de la variable del criterio, mientras que el modelo múltiple explica el 5,1% ($R^2$=0.051).

### 6.1.2. EJERCICIO PRÁCTICO 2

Los datos aparecen recopilados en las variables *Día1*, *Día2*, *Día3,* que indican el número de horas trabajadas en los tres días. Además, se han recopilado las variables *Sexo,* que toma valores 0 para mujeres y 1 para hombres, y *Edad*, que toma 1 para adolescentes, 2 para jóvenes y 3 para adultos.

Además, se incluyen tres datos relacionados con las terapias que ha recibido el sujeto. Todas ellas toman 1 si el sujeto ha recibido la terapia, y 0 en caso contrario. Las variables son *Estrés*, *Ansiedad* y *Fobia* (*terapia.sav*).

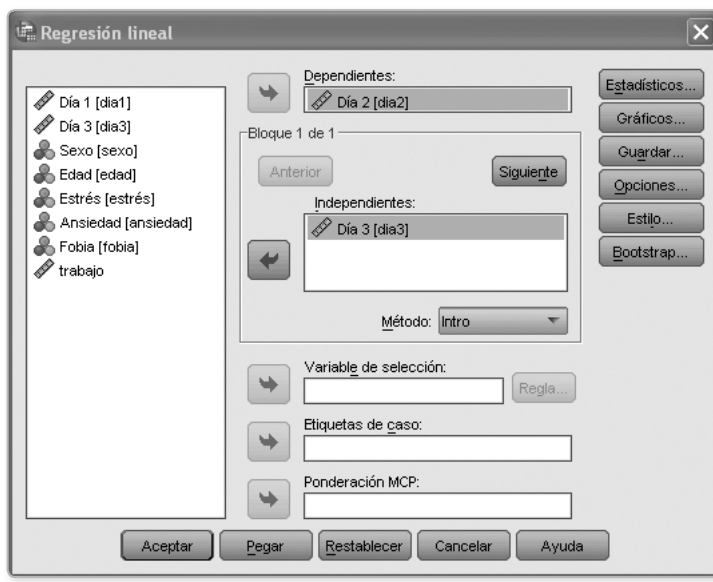
| Sujeto | Día1 | Día2 | Día3 | Sexo | Edad | Estrés | Ansiedad | Fobia |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.7 | 6.3 | 6.8 | 0 | 1 | 0 | 1 | 0 |
| 2 | 4.8 | 5.4 | 4.3 | 1 | 3 | 0 | 0 | 0 |
| 3 | 7.6 | 7.9 | 8.6 | 0 | 2 | 0 | 0 | 0 |
| 4 | 5.7 | 6.0 | 6.7 | 1 | 1 | 0 | 0 | 1 |
| 5 | 3.8 | 4.6 | 4.9 | 1 | 3 | 0 | 0 | 0 |
| 6 | 7.5 | 8.0 | 8.5 | 1 | 2 | 0 | 1 | 0 |
| 7 | 6.4 | 7.0 | 7.3 | 1 | 1 | 1 | 1 | 0 |
| 8 | 7.7 | 8.1 | 8.5 | 0 | 2 | 0 | 0 | 0 |
| 9 | 4.0 | 4.5 | 5.0 | 0 | 3 | 1 | 0 | 0 |
| 10 | 5.7 | 6.1 | 6.8 | 1 | 2 | 1 | 0 | 1 |

### A. Regresión

El procedimiento Analizar > Regresión lineal permite obtener la regresión lineal simple y múltiple. Por ejemplo, para estimar los parámetros del modelo:

$$Día2_i = \beta 0 + \beta_1 Día3_i + e_i$$

Se indica lo siguiente:




Los resultados muestran el coeficiente de determinación (R cuadrado) en la tabla Resumen del modelo. Se observa que el modelo estimado para *Día 2* explica el 89.5% de la varianza. La tabla de regresión en formato ANOVA muestra la descomposición de la suma de cuadrados total en la parte, debido a la regresión y el error:

**Resumen del modelo**

| Modelo | R | R cuadrado | R cuadrado ajustado | Error estándar de la estimación |
|---|---|---|---|---|
| 1 | ,946[a] | ,895 | ,882 | ,46019 |

a. Predictores: (Constante), Día 3

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 14.475 | 1 | 14.475 | 68.349 | .000[a] |
| | Residual | 1.694 | 8 | .212 | | |
| | Total | 16.169 | 9 | | | |

a. Predictors: (Constant), Día 3

b. Dependent Variable: Día 2

En el ejemplo, el nivel crítico asociado a F del ANOVA permite rechazar la hipótesis nula de que no existe relación lineal entre las variables ($H_0$: $\beta = 0$). A continuación, se muestran los valores estimados para los parámetros:

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | .965 | .672 | | 1.436 | .189 |
| | Día 3 | .805 | .097 | .946 | 8.267 | .000 |

a. Dependent Variable: Día 2

Estadístico de contraste T y nivel critico para $H_0$: $\beta = 0$.

La tabla de coeficientes muestra los términos de la ecuación de regresión. El término denominado «constante» es el origen de la ecuación (estimador de $\beta_0$). La pendiente $\beta_1$ (estimador de $\beta_1$) aparece en la fila de la tabla indicada por Día 3.

Los valores de $\beta_0$ y $\beta_1$ se muestran en la primera columna (Día2 = 0.965 + 0.805 Día3); a continuación, se presenta el error típico de estimación y los valores de $\beta_0$ y $\beta_1$, en caso que se utilicen puntuaciones típicas. Por último, se muestra el valor del estadístico "t" de contraste de las hipótesis nulas.

### 6.1.3. EJERCICIO PROPUESTO

1. ¿Existe relación entre el promedio de horas trabajas durante los tres primeras días y las horas dormidas un mes después de la terapia?
2. Calcule la regresión de la variable *Horas trabajadas transcurrido un mes después de la terapia* sobre la variable *Horas trabajadas durante la tercera noche*.
   a. ¿Cuál es la proporción de varianza en común entre las variables?
   b. Represente gráficamente la relación entre las variables.

3. El cuadro muestra las operaciones mensuales en una empresa de transporte de pasajeros, donde la variable dependiente está representada por los costos totales (miles), y la variable independiente por millas vehículo (miles). Encuentre el modelo y coeficiente de determinación.

| Mes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | 213.9 | 212.6 | 215.3 | 215.3 | 215.4 | 228.2 | 245.6 | 259.9 | 250.9 | 234.5 | 205.9 | 202.7 | 198.5 | 195.6 | 200.4 | 200.1 | 201.5 |
| X | 3147 | 3160 | 3197 | 3173 | 3292 | 3561 | 4013 | 4244 | 4159 | 3776 | 3232 | 3141 | 2928 | 3063 | 3096 | 3096 | 3158 |
| Mes | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | |
| Y | 213.2 | 219.5 | 243.7 | 262.3 | 252.3 | 224.4 | 215.3 | 202.5 | 200.7 | 201.8 | 202.1 | 200.4 | 209.3 | 213.9 | 227.0 | 246.4 | |
| X | 3338 | 3492 | 4019 | 4394 | 4251 | 3844 | 3276 | 3184 | 3037 | 3142 | 3159 | 3139 | 3203 | 3307 | 3585 | 4073 | |

   a. Calcule la regresión lineal de variable *Salario* frente a años de experiencia.

   b. Calcule el coeficiente de correlación lineal y el coeficiente de determinación.

   c. ¿Se observa alguna anomalía en el gráfico de los residuos frente a la variable *Regresora*?

   d. ¿El modelo es significativo?

4. La Dirección de una empresa quiere estudiar la rentabilidad de su inversión en publicidad. Para ello, ha recogido datos de volumen de ventas y del gasto en publicidad referidos a los años noventa y expresados en millones de soles.

| Año | Ventas | Gasto publicidad |
|---|---|---|
| 1990 | 50 | 10 |
| 1991 | 100 | 15 |
| 1992 | 150 | 18 |
| 1993 | 200 | 20 |
| 1994 | 200 | 25 |
| 1995 | 300 | 35 |
| 1996 | 400 | 50 |
| 1997 | 500 | 55 |
| 1998 | 650 | 60 |
| 1999 | 700 | 65 |

   a. Especifique y estime el modelo lineal que explique las ventas de la empresa, en función de la inversión publicitaria. Interprete los parámetros estimados.

   b. En el año 2004, la empresa va a invertir 550 000 unidades monetarias en publicidad. Calcule el volumen de ventas esperado.

   c. Se plantea el modelo $Yi = \beta Xi + Ei$; $i = 1... n$. Halle el estimador mínimo cuadrático de $\beta$.

   d. Aplique el resultado del apartado anterior para explicar el volumen de ventas, en función de los gastos en publicidad. Compruebe que la media de los residuos no es nula.

   f. Calcule la tabla ANOVA. ¿El modelo no es influyente?, ¿Por qué?

## 6.2. LOGLINEAL

El modelo loglineal permite resolver el problema que se enfrenta cotidianamente la investigación en el análisis de variables cualitativas (no numéricas). Debido a que las variables cualitativas no presentan un comportamiento lineal (ni reducible a lineal), al aplicar las técnicas convencionales de regresión, pueden generar modelos irreales o simplemente no convergen en algún valor.

**Regresión loglineal** es un método estadístico cuyo objetivo consiste en estudiar la "clasificación" de las variables cualitativas. Es esencialmente un modelo de regresión lineal múltiple entre las variables cualitativas y el logaritmo neperiano de la frecuencia de los datos (referenciales), de la siguiente forma:




El software estadístico SPSS permite determinar el modelo saturado y el modelo jerárquico más adecuado, mediante un algoritmo. De esta manera, se obtiene un modelo cuya variable de respuesta sea el logaritmo neperiano de la frecuencia ("conteo") de los datos (referenciales); y las variables de diseño, aquellas variables cualitativas seleccionadas para la determinación de valor de un inmueble. Las principales funciones se definen a continuación:

### A. General

Analiza las frecuencias de las observaciones incluidas en cada categoría de la clasificación cruzada de una tabla de contingencia. Cada una de las clasificaciones cruzadas de la tabla constituye una casilla y cada variable categórica se denomina «factor».

La ***variable dependiente*** es el número de casos (la frecuencia) en una casilla de la tabla de contingencia, y las ***variables explicativas*** son los factores y las covariables. Este procedimiento estima los parámetros de máxima verosimilitud de modelos loglineales jerárquicos y no jerárquicos, utilizando el método de Newton-Raphson. Es posible analizar una distribución multinomial o de Poisson.

Se pueden seleccionar hasta diez factores para definir las casillas de una tabla. Una variable de estructura de casilla permite definir ceros estructurales para tablas incompletas, incluir un término de desplazamiento en el modelo, ajustar un modelo log-tasa o implementar el método de corrección de las tablas marginales. Las ***variables de contraste*** permiten el cálculo del logaritmo de la razón de ventajas generalizadas (GLOR). SPSS muestra automáticamente información sobre el modelo y estadísticos de bondad de ajuste. Además, es posible mostrar una variedad de estadísticos y gráficos, o guardar los valores pronosticados y los residuos en el archivo de datos de trabajo.

### B. Logit

Analiza la relación entre variables dependientes (o de respuesta), y variables independientes (o explicativas). Las ***variables dependientes*** siempre son categóricas, mientras que las ***variables independientes*** pueden ser categóricas (factores). Otras variables independientes, las ***covariables de casilla***, pueden ser continuas, pero no se aplican en forma de caso por caso. A una casilla dada se le aplica la media ponderada de la covariable para los casos de esa casilla.

El logaritmo de las ventajas de las variables dependientes se expresa como una combinación lineal de parámetros. Se supone automáticamente una distribución multinomial; así, estos modelos se denominan a veces modelos logit multinomiales. Este procedimiento estima los parámetros de los modelos loglineales logit, utilizando el algoritmo de Newton-Raphson.

### C. Selección de modelo

Analiza las tablas de contingencia de varios factores. Ajusta modelos loglineales jerárquicos a las tablas de contingencia multidimensionales, usando un algoritmo de ajuste proporcional. Ayuda a encontrar cuáles de las variables categóricas están asociadas.

Para construir los modelos, se encuentran disponibles métodos de entrada forzada y de eliminación hacia atrás. Para los modelos saturados, es posible solicitar estimaciones de los parámetros y pruebas de asociación parcial. Un modelo saturado añade 0,5 a todas las casillas.

## 6.2.1. EJERCICIO PRÁCTICO 1

El ejercicio práctico consiste en realizar análisis de regresión por medio de los principales procedimientos de dicha herramienta, con los datos de la encuesta del supermercado.

### A. Modelos loglineales

Los modelos loglineales, a diferencia de los modelos de regresión múltiple, permiten un mejor análisis para el caso de variables cualitativas (categóricas) dicotómicas o politómicas.

$$\text{Ln(frecuencia)} = \mu + \lambda^{A} + \lambda^{B} + \lambda^{C} + \lambda^{AxB} + \lambda^{AxC} + \lambda^{BxC} + \lambda^{AxBxC}$$

El análisis loglineal se podría definir como un método, cuyo objetivo consiste en el estudio de las relaciones entre las variables cualitativas (o no numéricas). La regresión o análisis loglineal es un método estadístico cuyo objetivo consiste en estudiar la "clasificación" de las variables cualitativas. La regresión loglineal es esencialmente un modelo de regresión lineal múltiple entre las variables cualitativas y el logaritmo neperiano de la frecuencia de los datos (referenciales), de la forma:

A, B y C son variables cualitativas

Existen dos tipos de modelo:

#### a. Modelo saturado

Se define como modelo saturado (o completo) a aquel que contiene **todos** los posibles efectos principales, y **todas** las posibles combinaciones (efectos de segundo, tercer o enésimo orden) de la variables seleccionadas que lo componen. Debido a que el modelo loglineal saturado puede reproducir perfectamente la data estudiada, pues contiene todas las posibles combinaciones de las variables seleccionadas, se supone que un modelo pesado y complejo, usualmente no es el modelo más deseable.

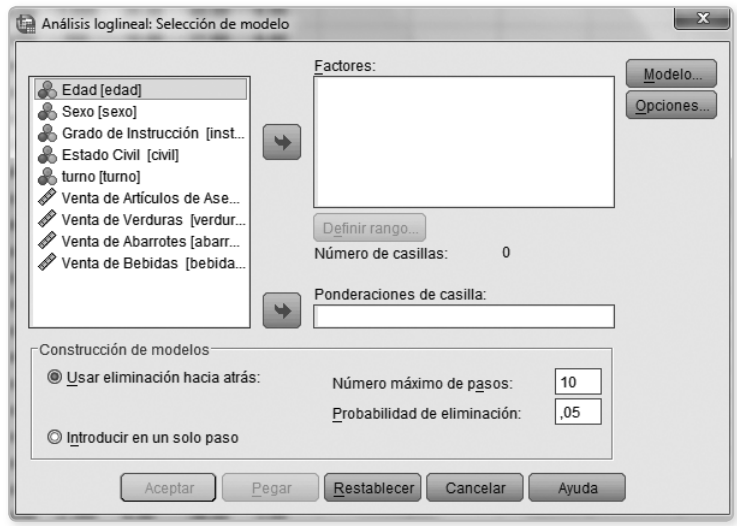
#### b. Modelos loglineal jerárquicos

Son los diferentes modelos, todos subjuegos (ecuaciones de menor orden que el modelo saturado) provenientes del modelo loglineal saturado que cumplan las condiciones siguientes:

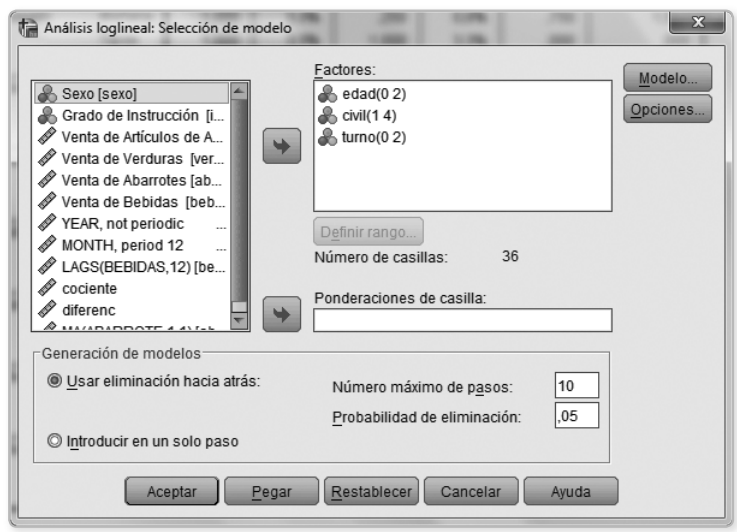
- Si un parámetro es nulo, también los serán aquellos términos de orden inferior.
- Que exista completa independencia entre las variables seleccionadas.
- Si estas condiciones se cumplen, se genera un modelo loglineal más sencillo, elegante y con un grado aceptable de precisión.

Se presenta el siguiente caso práctico, utilizando el paquete estadístico SPSS. Los procedimientos de cálculo en el SPSS se encuentran en el menú Analyze > Loglineal. Dentro del módulo loglineal, se encuentran los submenús: General..., Logit... y Model Selection...

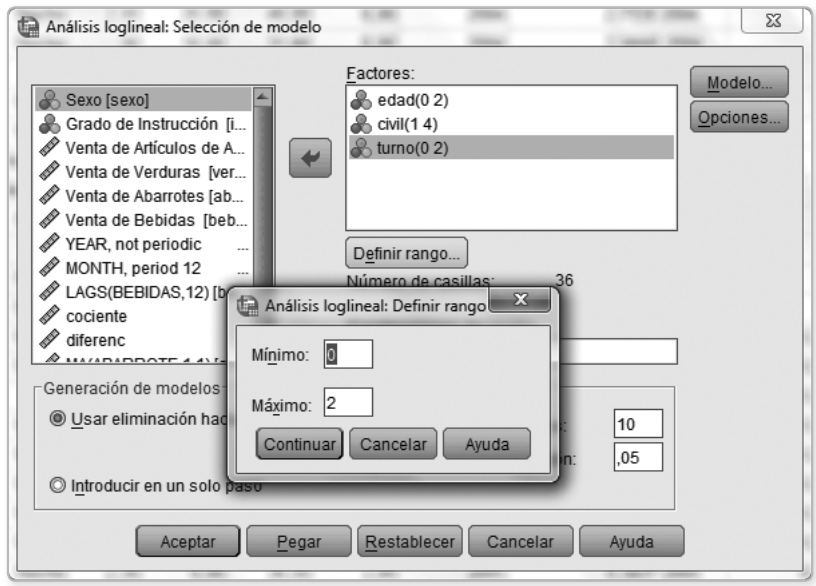
Si se seleccionan las variables categóricas (edad, estado civil, grado de instrucción, turno) y utilizando el procedimiento **Análisis > Loglineal > Selección de modelo**, se tiene el cuadro de diálogo siguiente:




De la lista de variables del cuadro de la izquierda, se seleccionan las variables *Edad, Estado civil* y *Turno.* Luego se pasa al cuadro del lado derecho como se observa en el siguiente gráfico:



Enseguida, presione Definir rango (Define Range...) donde tiene que escribir el rango mínimo y rango máximo para cada variable; por ejemplo, *Edad* tiene rango mínimo igual 0 y rango máximo igual 2, *Estado civil* tiene rango mínimo igual 1 y rango máximo igual 4, y *Turno* tiene rango mínimo igual 0 y rango máximo igual 2.




En el cuadro Definir rangos, presione el botón Continuar y luego Aceptar (OK), se obtiene los siguientes resultados:

**Análisis de loglinear jerárquico**

**Información de datos**

| | | N |
|---|---|---|
| Casos | Válido | 30 |
| | Fuera de rango[a] | 0 |
| | Perdidos | 0 |
| | Ponderado válido | 30 |
| Categorías | Edad | 3 |
| | Estado Civil | 4 |
| | turno | 3 |

a. Casos rechazados debido a valores de factor fuera de rango.

**Diseño 1**

**Información de convergencia**

| | |
|---|---|
| Clase generadora | edad*civil*turno |
| Número de iteraciones | 1 |
| Diferencia máxima entre marginales observados y ajustados | ,000 |
| Criterio de convergencia | ,250 |

**Recuentos de casilla y residuos**

| Edad | Estado Civil | turno | Observado | | Esperado | | Residuos | Residuos estándar |
|---|---|---|---|---|---|---|---|---|
| | | | Recuento[a] | % | Recuento | % | | |
| Adolescente | Soltero | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | Casado | Mañana | 3,500 | 11,7% | 3,500 | 11,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | Divorciado | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | Viudo | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| Joven | Soltero | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | 2,500 | 8,3% | 2,500 | 8,3% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | Casado | Mañana | 1,500 | 5,0% | 1,500 | 5,0% | ,000 | ,000 |
| | | Tarde | 1,500 | 5,0% | 1,500 | 5,0% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | Divorciado | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | Viudo | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | 1,500 | 5,0% | 1,500 | 5,0% | ,000 | ,000 |
| | | Noche | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| Adulto | Soltero | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | 3,500 | 11,7% | 3,500 | 11,7% | ,000 | ,000 |
| | Casado | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | 7,500 | 25,0% | 7,500 | 25,0% | ,000 | ,000 |
| | Divorciado | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | 4,500 | 15,0% | 4,500 | 15,0% | ,000 | ,000 |
| | Viudo | Mañana | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Tarde | ,500 | 1,7% | ,500 | 1,7% | ,000 | ,000 |
| | | Noche | 8,500 | 28,3% | 8,500 | 28,3% | ,000 | ,000 |

a. Para los modelos saturados, ,500 se ha añadido a todas las casillas observadas.

**Pruebas de bondad de ajuste**

| | Chi-cuadrado | gl | Sig. |
|---|---|---|---|
| Razón de verosimilitud | ,000 | 0 | . |
| Pearson | ,000 | 0 | . |

**Efectos de K y de orden superior**

| | K | gl | Razón de verosimilitud | | Pearson | | Número de iteraciones |
|---|---|---|---|---|---|---|---|
| | | | Chi-cuadrado | Sig. | Chi-cuadrado | Sig. | |
| Efectos K y de orden superior[a] | 1 | 35 | 98,499 | ,000 | 154,800 | ,000 | 0 |
| | 2 | 28 | 52,453 | ,003 | 106,351 | ,000 | 2 |
| | 3 | 12 | ,000 | 1,000 | ,000 | 1,000 | 3 |
| Efectos K[b] | 1 | 7 | 46,047 | ,000 | 48,449 | ,000 | 0 |
| | 2 | 16 | 52,453 | ,000 | 106,351 | ,000 | 0 |
| | 3 | 12 | ,000 | 1,000 | ,000 | 1,000 | 0 |

El gl utilizado para estas pruebas NO se ha ajustado para los ceros estructurales o de muestreo. Las pruebas que utilizan este gl podrían ser conservadoras.

a. Prueba que los efectos K y de orden superior son cero.

b. Prueba que los efectos K son cero.

## Estadísticas de eliminación hacia atrás

**Resumen de los pasos**

| Escalón[a] | | | Effects | Chi-cuadrado[c] | gl | Sig. | Número de iteraciones |
|---|---|---|---|---|---|---|---|
| 0 | Clase generadora[b] | | edad*civil*turno | ,000 | 0 | . | |
| | Efecto suprimido | 1 | edad*civil*turno | ,000 | 12 | 1,000 | 3 |
| 1 | Clase generadora[b] | | edad*civil, edad*turno, civil*turno | ,000 | 12 | 1,000 | |
| | Efecto suprimido | 1 | edad*civil | ,000 | 6 | 1,000 | 2 |
| 2 | Clase generadora[b] | | edad*turno, civil*turno | ,000 | 18 | 1,000 | |
| | Efecto suprimido | 1 | edad*turno | 40,881 | 4 | ,000 | 2 |
| | | 2 | civil*turno | 11,572 | 6 | ,072 | 2 |
| 3 | Clase generadora[b] | | edad*turno, civil | 11,572 | 24 | ,984 | |
| | Efecto suprimido | 1 | edad*turno | 40,881 | 4 | ,000 | 2 |
| | | 2 | civil | 5,478 | 3 | ,140 | 2 |
| 4 | Clase generadora[b] | | edad*turno | 17,050 | 27 | ,930 | |
| | Efecto suprimido | 1 | edad*turno | 40,881 | 4 | ,000 | 2 |
| 5 | Clase generadora[b] | | edad*turno | 17,050 | 27 | ,930 | |

a. En cada paso, se suprime el efecto con el nivel de significación más grande para el cambio de la razón de verosimilitud, siempre que el nivel de significación sea mayor que ,050.

b. Las estadísticas se visualizan para el mejor modelo en cada paso después del paso 0.

c. Para 'Efecto suprimido', este es el cambio en el chi-cuadrado después de que se suprima el efecto del modelo.

**Información de convergencia[a]**

| | |
|---|---|
| Clase generadora | edad*turno |
| Número de iteraciones | 0 |
| Diferencia máxima entre marginales observados y ajustados | ,000 |
| Criterio de convergencia | ,250 |

**Pruebas de bondad de ajuste**

| | Chi-cuadrado | gl | Sig. |
|---|---|---|---|
| Razón de verosimilitud | 17,050 | 27 | ,930 |
| Pearson | 17,091 | 27 | ,929 |

**Recuentos de casilla y residuos**

| Edad | Estado Civil | turno | Observado | | Esperado | | Residuos | Residuos estándar |
|---|---|---|---|---|---|---|---|---|
| | | | Recuento | % | Recuento | % | | |
| Adolescente | Soltero | Mañana | ,000 | 0,0% | ,750 | 2,5% | -,750 | -,866 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | Casado | Mañana | 3,000 | 10,0% | ,750 | 2,5% | 2,250 | 2,598 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | Divorciado | Mañana | ,000 | 0,0% | ,750 | 2,5% | -,750 | -,866 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | Viudo | Mañana | ,000 | 0,0% | ,750 | 2,5% | -,750 | -,866 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| Joven | Soltero | Mañana | ,000 | 0,0% | ,250 | 0,8% | -,250 | -,500 |
| | | Tarde | 2,000 | 6,7% | 1,000 | 3,3% | 1,000 | 1,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | Casado | Mañana | 1,000 | 3,3% | ,250 | 0,8% | ,750 | 1,500 |
| | | Tarde | 1,000 | 3,3% | 1,000 | 3,3% | ,000 | ,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | Divorciado | Mañana | ,000 | 0,0% | ,250 | 0,8% | -,250 | -,500 |
| | | Tarde | ,000 | 0,0% | 1,000 | 3,3% | -1,000 | -1,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | Viudo | Mañana | ,000 | 0,0% | ,250 | 0,8% | -,250 | -,500 |
| | | Tarde | 1,000 | 3,3% | 1,000 | 3,3% | ,000 | ,000 |
| | | Noche | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| Adulto | Soltero | Mañana | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | 3,000 | 10,0% | 5,500 | 18,3% | -2,500 | -1,066 |
| | Casado | Mañana | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | 7,000 | 23,3% | 5,500 | 18,3% | 1,500 | ,640 |
| | Divorciado | Mañana | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | 4,000 | 13,3% | 5,500 | 18,3% | -1,500 | -,640 |
| | Viudo | Mañana | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Tarde | ,000 | 0,0% | ,000 | 0,0% | ,000 | ,000 |
| | | Noche | 8,000 | 26,7% | 5,500 | 18,3% | 2,500 | 1,066 |

CAPÍTULO

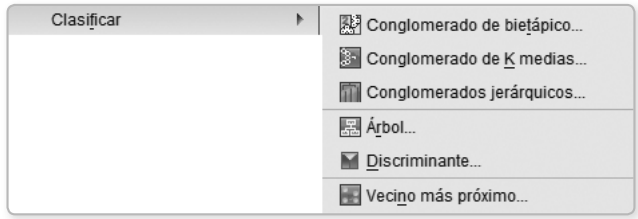
7

# CLASIFICAR (CLÚSTER) Y REDUCCIÓN DE DATOS

Los temas que se presentan en este capítulo son: conglomerados de K medias, conglomerados jerárquicos, análisis discriminante, análisis de componentes principales, análisis factorial y análisis de correspondencia.

## 7.1. CLASIFICAR

Es una técnica estadística que permite el agrupamiento de variables; además, permite la agrupación de casos. SPSS dispone los siguientes procedimientos de análisis de *conglomerados de K medias*, el de *conglomerados jerárquicos* y el análisis *discriminante*.




### 7.1.1. CONGLOMERADOS DE K MEDIAS

Intenta identificar grupos de casos relativamente homogéneos, basándose en las características seleccionadas y utilizando un algoritmo que puede gestionar un gran número de casos. Sin embargo, el algoritmo requiere que el usuario especifique el número de conglomerados.

Se pueden especificar los centros iniciales de los conglomerados si se conoce de antemano dicha información; además, se puede elegir uno de los dos métodos disponibles para clasificar los casos: la actualización de los centros de los conglomerados de forma iterativa, o solo la clasificación.

Asimismo, se puede guardar la pertenencia a los conglomerados, la información de la distancia y los centros de los conglomerados finales. Permite especificar una variable cuyos valores sean utilizados para etiquetar los resultados por casos.

También, se pueden solicitar los estadísticos F de los análisis de varianza. Aunque estos estadísticos son oportunistas (ya que el procedimiento trata de formar grupos que de hecho difieran), el tamaño relativo de los estadísticos proporciona información acerca de la contribución de cada variable a la separación de los grupos.

### 7.1.2. CONGLOMERADOS JERÁRQUICOS

Intenta identificar grupos relativamente homogéneos de casos (o de variables), basándose en las características seleccionadas, mediante un algoritmo que comienza con cada caso (o cada variable) en un conglomerado diferente, y combina los conglomerados hasta que solo queda uno.

Es posible analizar las variables brutas o elegir entre una variedad de transformaciones de estandarización. Las medidas de distancia o similaridad se generan mediante el procedimiento *Proximidades*. Los estadísticos se muestran en cada etapa para ayudar a seleccionar la mejor solución.

### 7.1.3. DISCRIMINANTE

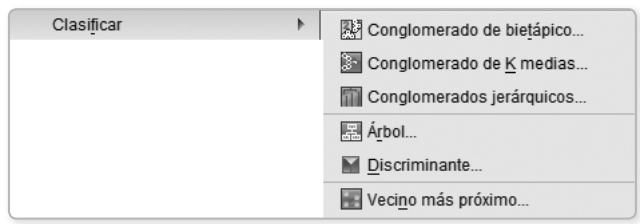
Resulta útil para las situaciones en las que se desea construir un modelo predictivo, para pronosticar el grupo de pertenencia de un caso a partir de las características observadas de cada caso.

El procedimiento genera una función discriminante (o para más de dos grupos, un conjunto de funciones discriminantes), basada en combinaciones lineales de las variables predictoras, que proporcionan la mejor discriminación posible entre los grupos.

Las funciones se generan a partir de una muestra de casos, para los que se conoce el grupo de pertenencia; posteriormente, las funciones pueden ser aplicadas a nuevos casos que dispongan de medidas para las variables predictoras, pero de los que se desconozca el grupo de pertenencia. La variable de agrupación puede tener más de dos valores, y sus códigos han de ser números enteros, siendo necesario especificar sus valores máximo y mínimo. Los casos con valores fuera de estos límites se excluyen del análisis.

## EJERCICIO PRÁCTICO 1

El ejercicio práctico permitirá usar los principales procedimientos del menú Clasificar. Los datos usados serán los recopilados en la encuesta del supermercado. Dentro del menú Analizar, se encuentra el procedimiento **Clasificar,** que permite agrupar variables y casos. El aspecto de dicho menú es el siguiente **Analizar** > **Clasificar** (Classity).




Para realizar el procedimiento Clasificar, se deben seguir los cuadros de diálogo que se presenta. A continuación, se muestra una explicación detallada del cuadro de diálogo correspondiente a los principales procedimientos.

### A. Conglomerados de K medias: Fundamentos del algoritmo de K-medias

Una muestra de n elementos con p variables se divide en un número de grupos prefijado, K. El algoritmo de K-medias requiere cuatro etapas:

1. Seleccione K puntos como centros de los grupos iniciales, de la siguiente manera:
   a. Asigne aleatoriamente los objetos a los grupos, tomando los centros de los grupos así formados.
   b. Tome como centros los K puntos más alejados entre sí.
   c. Construya los grupos con información a priori o bien seleccionando los centros a priori.
2. Calcule las distancias euclídeas de cada elemento al centro de los K grupos y asigne cada elemento al grupo más próximo. La asignación se realiza secuencialmente, y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.
3. Defina un criterio de optimalidad y compruebe si reasignando uno a uno cada elemento de un grupo a otro, mejora el criterio.
4. Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

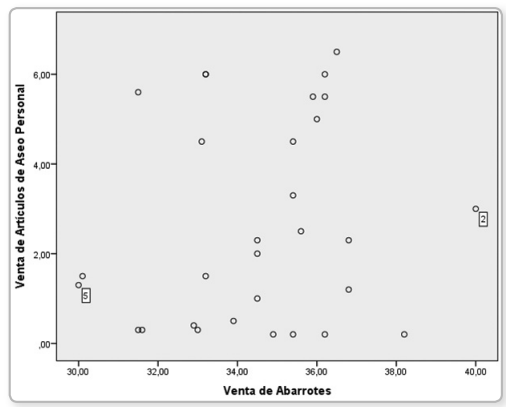
El **análisis de conglomerados** (*cluster analisys*) es una técnica multivariante que permite agrupar los casos o variables de un archivo de datos, en función del parecido o similaridad existente entre ellos.

Como técnica de agrupación de variables, el análisis de conglomerados es similar al análisis factorial; sin embargo, mientras que la factorización es más bien poco flexible en algunos de sus supuestos (linealidad, normalidad, variables cuantitativas, etc.) y siempre estima, de la misma manera, la matriz de distancias, la aglomeración es menos restrictiva en sus supuestos (no exige linealidad ni simetría, permite variables categóricas, etc.) y admite varios métodos de estimación de la matriz de distancias.

Como técnica de agrupación de casos, el análisis de conglomerados es similar al análisis discriminante. Sin embargo, mientras que el análisis discriminante efectúa la clasificación, tomando como referencia un criterio o variable dependiente (los grupos de clasificación), el análisis de conglomerados permite detectar el número óptimo de grupos y su composición, únicamente a partir de la similaridad existente entre los casos; además, el análisis de conglomerados no asume alguna distribución específica para las variables. Por simplicidad, en este capítulo se comienza exponiendo la agrupación de casos.

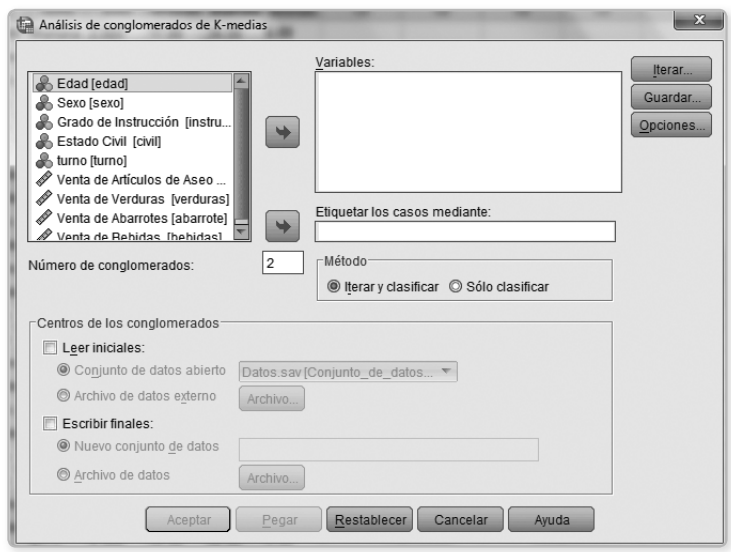
- **Gráfico de dispersión**

Para llevar a cabo un análisis de conglomerados de K medias, se recomienda realizar un gráfico de dispersión, con el fin de tener una primera lectura acerca de los datos.

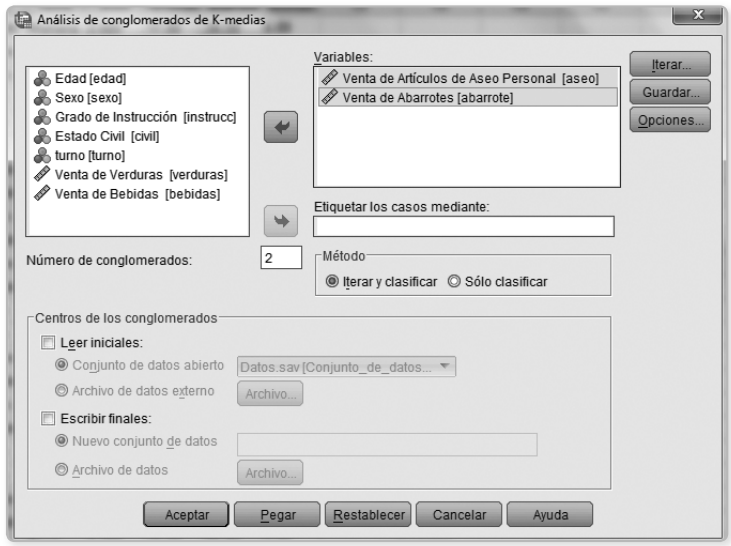



Del gráfico se observan los centros iniciales que están representados por el caso 5 y el caso 2; es decir, son los valores que corresponden a los dos casos que han sido elegidos como centros respectivos de los conglomerados, en las dos variables de clasificación.

A continuación, seleccione la opción **Analizar > Clasificar > Conglomerados de K medias**, obteniéndose el cuadro de diálogo siguiente:



El cuadro de la izquierda muestra la lista de variables del archivo (numérico y cadena); pero las variables de cadena solo sirven para etiquetar casos. Para obtener un análisis de conglomerados de K medias, se seleccionan las variables con los que se quieren formar los grupos y se traslada al cuadro del lado derecho. En este caso, las variables son *Venta de artículos de aseo personal* y *Venta de abarrotes*.




Opcionalmente, se selecciona una variable para identificar los casos en las tablas de resultados y en los gráficos. Luego, traslade a **Etiquetar casos mediante** (Label Cases by).

El **Número de conglomerados** (*Number of Clusters*) se encuentra seleccionado, por defecto, dos conglomerados. Para solicitar un número mayor de conglomerados, introduzca el número deseado en el cuadro. Existen dos métodos (Method): el primero, **Iterar y clasificar** (*Iterar and classify*) es un procedimiento que se encarga de estimar los centros iterativamente y de clasificar a los sujetos con arreglo a los centros estimados; el segundo, **Solo clasificar** (*classify only*) se clasifica a los sujetos, según los centros iniciales (sin actualizar sus valores iterativamente). Este botón se suele utilizar junto con el botón **Centros>>**. Luego de realizar estas selecciones se presiona Aceptar (OK), y el visor presenta los siguientes resultados: el primer cuadro muestra los centros iniciales, es decir, los valores que corresponden a los dos casos que han sido elegidos como centros respectivos de los conglomerados solicitados, en las dos variables de clasificación utilizadas.

**Initial Cluster Centers**

| | Cluster | |
|---|---|---|
| | 1 | 2 |
| Venta de Artículos de Aseo Personal | 1.30 | 3.00 |
| Venta de Abarrotes | 30.00 | 40.00 |

Se observa que los valores 1.30 y 3.00 son los centros iniciales del primer y segundo grupo, respectivamente, de la variable *Venta de artículos de aseo personal*. Por otro lado, los valores 30.00 y 40.00 representan los centros iniciales del primer y segundo grupo de la variable *Venta abarrotes*, tal como se visualiza en el gráfico de dispersión.

Una vez seleccionados los centros de los conglomerados, cada caso es asignado al conglomerado cuyo centro se encuentra más próximo, y comienza con un proceso de ubicación iterativa de los centros.

En la primera iteración, se reasignan los casos por su distancia al nuevo centro y, tras la reasignación, se vuelve a actualizar el valor del centro. En la siguiente iteración, se vuelven a reasignar los casos y a actualizar el valor del centro.

El cuadro siguiente resume el historial de iteraciones (3, en este ejemplo) con indicación del cambio (desplazamiento) experimentado por cada centro en cada iteración. Puede observarse que, conforme avanzan las iteraciones, el desplazamiento de los centros se va haciendo más y más pequeño, hasta llegar a la tercera iteración, en la que ya no existe desplazamiento alguno.

**Iteration History[a]**

| | Change in Cluster Centers | |
|---|---|---|
| Iteration | 1 | 2 |
| 1 | 2.962 | 3.540 |
| 2 | .187 | .251 |
| 3 | .000 | .000 |

a. Convergence achieved due to no or small distance change. The maximum distance by which any center has changed is .000. The current iteration is 3. The minimum distance between initial centers is 10.143.

En el SPSS, el proceso de iteración se detiene por defecto cuando alcanza 10 (u otro número ingresado), o cuando de una iteración a otra no se produce algún cambio. En este ejemplo, se detiene en la tercera iteración porque no se produce cambio alguno.

El cuadro siguiente ofrece los centros de los conglomerados finales, es decir, los centros de los conglomerados tras el proceso de actualización iterativa. Se comparan los centros finales (tras la iteración) de esta tabla con los centros iniciales (antes de la iteración).

**Final Cluster Centers**

| | Cluster | |
|---|---|---|
| | 1 | 2 |
| Venta de Artículos de Aseo Personal | 1.99 | 3.52 |
| Venta de Abarrotes | 33.00 | 36.55 |

Se puede apreciar que el primer (1) conglomerado está constituido por las ventas en menor cantidad en nuevos soles, tanto en artículos de aseo personal y de abarrotes y el segundo conglomerado (2) por las ventas en mayor cantidad en nuevos soles de aseo personal y de abarrotes.

El cuadro siguiente muestra el número de casos asignados a cada conglomerado.

**Number of Cases in each Cluster**

| | | |
|---|---|---|
| Cluster | 1 | 17.000 |
| | 2 | 13.000 |
| Valid | | 30.000 |
| Missing | | .000 |

Respecto a los tamaños de los conglomerados, el primero tiene 17 y el segundo, 13.

El cuadro siguiente es la tabla de ANOVA, que muestra un resumen del análisis de varianza con una estadística F univariada para cada una de las variables incluidas en el análisis.

**ANOVA**

| | Cluster | | Error | | | |
|---|---|---|---|---|---|---|
| | Mean Square | df | Mean Square | df | F | Sig. |
| Venta de Artículos de Aseo Personal | 17.048 | 1 | 4.627 | 28 | 3.685 | .065 |
| Venta de Abarrotes | 93.040 | 1 | 2.155 | 28 | 43.179 | .000 |

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

El cuadro siguiente muestra un listado de todos los casos usados en el análisis, con indicación del conglomerado al que ha sido asignado cada caso y la distancia euclídea existente entre cada caso y el centro de su conglomerado.

**Cluster Membership**

| Case Number | Cluster | Distance |
|---|---|---|
| 1 | 1 | 1.500 |
| 2 | 2 | 3.484 |
| 3 | 1 | 2.198 |
| 4 | 2 | 1.174 |
| 5 | 1 | 3.079 |
| 6 | 1 | 1.597 |
| 7 | 1 | .533 |
| 8 | 1 | 2.508 |
| 9 | 2 | 1.393 |
| 10 | 1 | 1.694 |
| 11 | 1 | 1.800 |
| 12 | 1 | 4.011 |
| 13 | 1 | 3.905 |
| 14 | 2 | 2.510 |
| 15 | 2 | 2.328 |
| 16 | 1 | 2.996 |
| 17 | 1 | 4.011 |
| 18 | 2 | 2.090 |
| 19 | 2 | 2.985 |
| 20 | 2 | 3.702 |
| 21 | 1 | 1.531 |
| 22 | 2 | 3.334 |
| 23 | 2 | 1.240 |
| 24 | 1 | 2.942 |
| 25 | 2 | 1.585 |
| 26 | 2 | 1.517 |
| 27 | 1 | 2.263 |
| 28 | 2 | 2.016 |
| 29 | 1 | 1.744 |
| 30 | 1 | 2.613 |

El siguiente cuadro muestra la distancia euclídea existente entre los centros de los conglomerados finales.

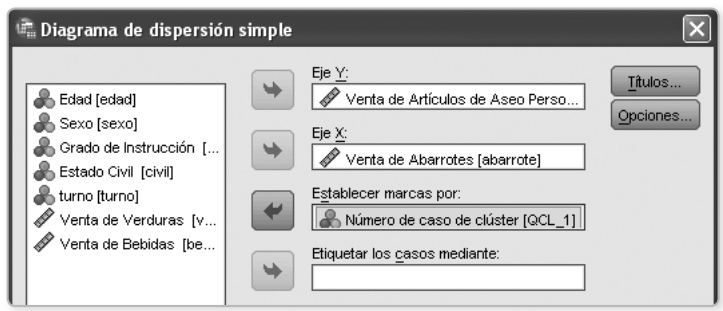
**Number of Cases in each Cluster**

| | | |
|---|---|---|
| Cluster | 1 | 17.000 |
| | 2 | 13.000 |
| Valid | | 30.000 |
| Missing | | .000 |

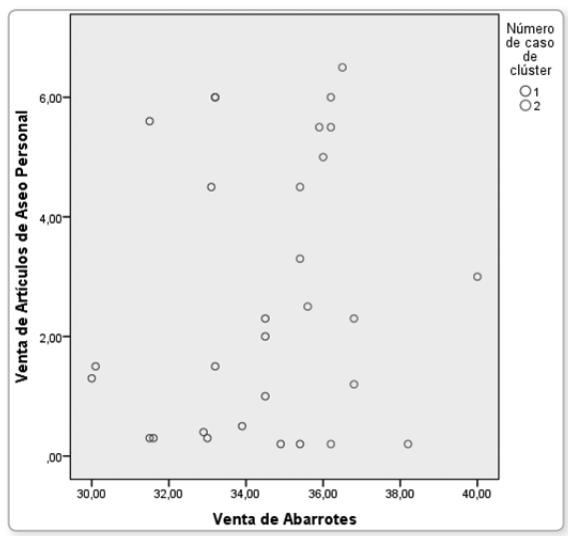
Además, se pueden guardar los clústers de pertenencia; es decir, en qué grupo han sido clasificados los casos. Para ello, haga clic en Guardar, luego seleccione Clúster de pertenencia y la clasificación aparecerá en la Vista de datos.




A continuación, se puede realizar un gráfico de dispersión con el fin de visualizar cómo quedan clasificados los grupos finales. Para ello, elija el gráfico de dispersión. El cuadro de diálogo es el siguiente:




Luego, se obtiene el siguiente gráfico que representa la clasificación final de los grupos (clúster).



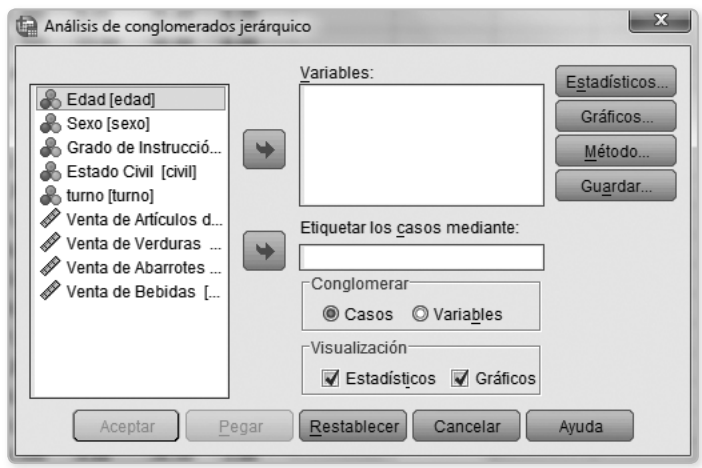

Se observan los casos divididos en dos grupos finales.

### B. Conglomerados jerárquico

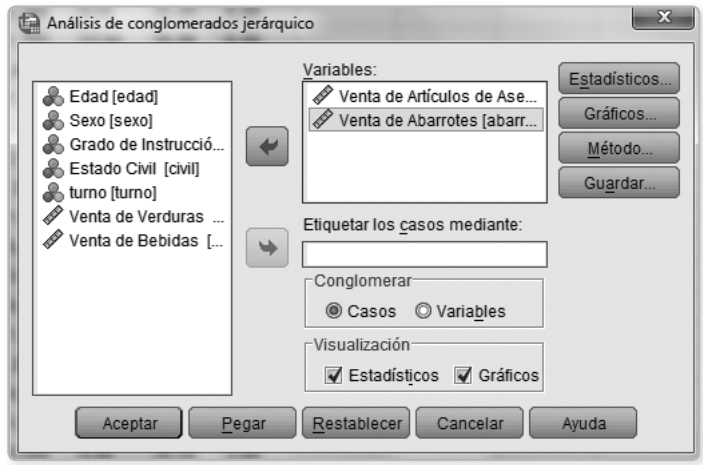
A diferencia de lo que ocurre con el procedimiento Análisis de conglomerados de K medias, el procedimiento Análisis de conglomerados jerárquico permite aglomerar tanto casos como variables y elegir entre una gran variedad de métodos de aglomeración y medidas de distancia. Pero la diferencia fundamental entre ambos procedimientos está en que, en el segundo de ellos, se procede de forma jerárquica.

El análisis de conglomerados jerárquico comienza con el cálculo de la matriz de distancias entre los elementos de la muestra (casos o variables). Esa matriz contiene las distancias existentes entre cada elemento y todos los restantes de la muestra. A continuación, se buscan los dos elementos más próximos (es decir, los dos más similares en términos de distancia) y se agrupan en un conglomerado. El conglomerado resultante es indivisible a partir de ese momento: de ahí el nombre de «jerárquico» asignado al procedimiento. De esta manera, se van agrupando los elementos en conglomerados cada vez más grandes y más heterogéneos hasta llegar al último paso, en el que todos los elementos muestrales quedan agrupados en un único conglomerado global. En cada paso del proceso, pueden agruparse casos individuales, conglomerados previamente formados o un caso individual con un conglomerado previamente formado. El **análisis de conglomerados jerárquico** es, por tanto, una técnica aglomerativa: partiendo de los elementos muestrales individualmente considerados, va creando grupos hasta llegar a la formación de un único grupo o conglomerado, constituido por todos los elementos de la muestra.

Para llevar a cabo un análisis de conglomerados de K medias, se selecciona la opción Analizar > Clasificar > Conglomerados jerárquicos, obteniéndose el cuadro de diálogo siguiente:




El cuadro de la izquierda muestra la lista de variables del archivo (numérico y cadena). Pero las variables de cadena solo sirven para etiquetar casos. Para obtener un análisis de conglomerados jerárquicos, se seleccionan las variables numéricas que se desea utilizar para diferenciar los casos y formar conglomerados; además, se traslada al cuadro del lado derecho. En este caso, las variables son Venta de artículos de aseo personal y Venta de abarrotes.




Opcionalmente, se selecciona una variable para identificar los casos en las tablas de resultado y en los gráficos. Luego traslade a Etiquetar casos mediante (Label Cases by).

- **Conglomerar (Custers).** Permite decidir qué elementos del archivo de datos se desea agrupar.
- **Casos (Cases)**. Se agrupan los casos a partir de sus puntuaciones en las variables seleccionadas, y es la opción por defecto.
- **Variables (Variables)**. Se agrupan las variables seleccionadas en la lista Variables a partir de las puntuaciones de los casos válidos del archivo de datos. Esta opción exige incluir en el análisis al menos tres variables.
- **Mostrar (Display)**. Las opciones permiten controlar el tipo de resultado que mostrará el Visor de resultados, seleccionadas por defecto.
  - **Estadísticas (Statistics).** El Visor de resultados solo muestra las tablas de resultados. Desactivando esta opción, se anula el botón Estadísticas.
  - **Gráficos (Plots).** El Visor de resultados solo muestra los gráficos.

### Clasificar por casos

Luego de realizar estas selecciones, presione Aceptar (OK) y el visor presentará los siguientes resultados. El primer cuadro muestra un resumen de los casos procesados, el número y porcentaje de casos válidos analizados, el número y porcentaje con valores perdidos, y el tamaño total de la muestra.

Resumen de procesamiento de casos[a,b]

| Casos | | | | | |
|---|---|---|---|---|---|
| Válido | | Perdidos | | Total | |
| N | Porcentaje | N | Porcentaje | N | Porcentaje |
| 30 | 100,0 | 0 | ,0 | 30 | 100,0 |

a. Distancia euclídea al cuadrado utilizada
b. Enlace promedio (entre grupos)

**NOTA**

*"a" indica el nombre de la medida utilizada para obtener la matriz de distancia Squared Euclidean Distance used (Distancia euclídea al cuadrado). "b" el método de conglomerados utilizando Vinculación promedio (Intergrupos).*

El cuadro siguiente muestra el historial del proceso de conglomerados, etapa por etapa. En cada etapa se unen dos elementos. Como la muestra analizada tiene 30 casos, solo se realizan 29 etapas de fusión.

La columna *Conglomerados que se combina* (Cluster Combinet) informa sobre los conglomerados (o casos) fundidos en cada etapa. En este ejemplo, en la primera etapa se han fundido 12 y 17.

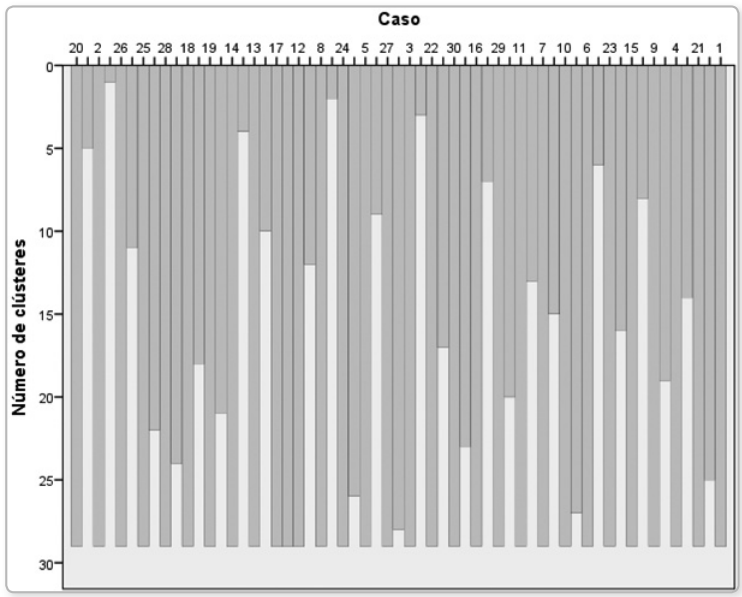
**Historial de conglomeración**

| Etapa | Clúster combinado | | Coeficientes | Primera aparición del clúster de etapa | | Etapa siguiente |
|---|---|---|---|---|---|---|
| | Clúster 1 | Clúster 2 | | Clúster 1 | Clúster 2 | |
| 1 | 12 | 17 | ,000 | 0 | 0 | 18 |
| 2 | 3 | 27 | ,010 | 0 | 0 | 21 |
| 3 | 6 | 10 | ,020 | 0 | 0 | 15 |
| 4 | 5 | 24 | ,050 | 0 | 0 | 21 |
| 5 | 1 | 21 | ,090 | 0 | 0 | 16 |
| 6 | 18 | 28 | ,090 | 0 | 0 | 8 |
| 7 | 16 | 30 | ,250 | 0 | 0 | 13 |
| 8 | 18 | 25 | ,275 | 6 | 0 | 12 |
| 9 | 14 | 19 | ,340 | 0 | 0 | 12 |
| 10 | 11 | 29 | ,610 | 0 | 0 | 17 |
| 11 | 4 | 9 | ,680 | 0 | 0 | 16 |
| 12 | 14 | 18 | 1,097 | 9 | 8 | 19 |
| 13 | 16 | 22 | 1,165 | 7 | 0 | 23 |
| 14 | 15 | 23 | 1,210 | 0 | 0 | 22 |
| 15 | 6 | 7 | 1,390 | 3 | 0 | 17 |
| 16 | 1 | 4 | 1,755 | 5 | 11 | 22 |
| 17 | 6 | 11 | 1,825 | 15 | 10 | 23 |
| 18 | 8 | 12 | 2,260 | 0 | 1 | 20 |
| 19 | 14 | 26 | 2,320 | 12 | 0 | 26 |
| 20 | 8 | 13 | 3,290 | 18 | 0 | 26 |
| 21 | 3 | 5 | 3,475 | 2 | 4 | 27 |
| 22 | 1 | 15 | 4,630 | 16 | 14 | 24 |
| 23 | 6 | 16 | 5,153 | 17 | 13 | 24 |
| 24 | 1 | 6 | 7,564 | 22 | 23 | 27 |
| 25 | 2 | 20 | 11,080 | 0 | 0 | 29 |
| 26 | 8 | 14 | 12,212 | 20 | 19 | 28 |
| 27 | 1 | 3 | 19,854 | 24 | 21 | 28 |
| 28 | 1 | 8 | 27,633 | 27 | 26 | 29 |
| 29 | 1 | 2 | 36,753 | 28 | 25 | 0 |

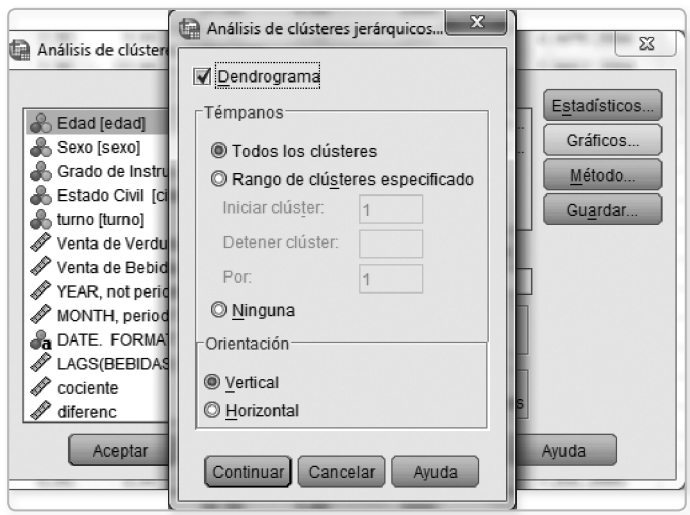
La columna *Coeficientes* (coefficients) es el valor de la distancia a la que se encuentran los casos antes de fusión. En la primera etapa, la distancia de fusión entre los casos 12 y 17 vale 0.000.

La columna *Etapa en la que el conglomerado aparece por primera vez (stage cluster First Appears)* indica la etapa donde se han formado los conglomerados que se están fundiendo en cada momento. Por ejemplo, «Cero» indica que el conglomerado es un caso individual. La columna *Próxima* etapa (Next Stage) indica la etapa donde el conglomerado que se acaba de formar volverá a fundirse con otros elementos.

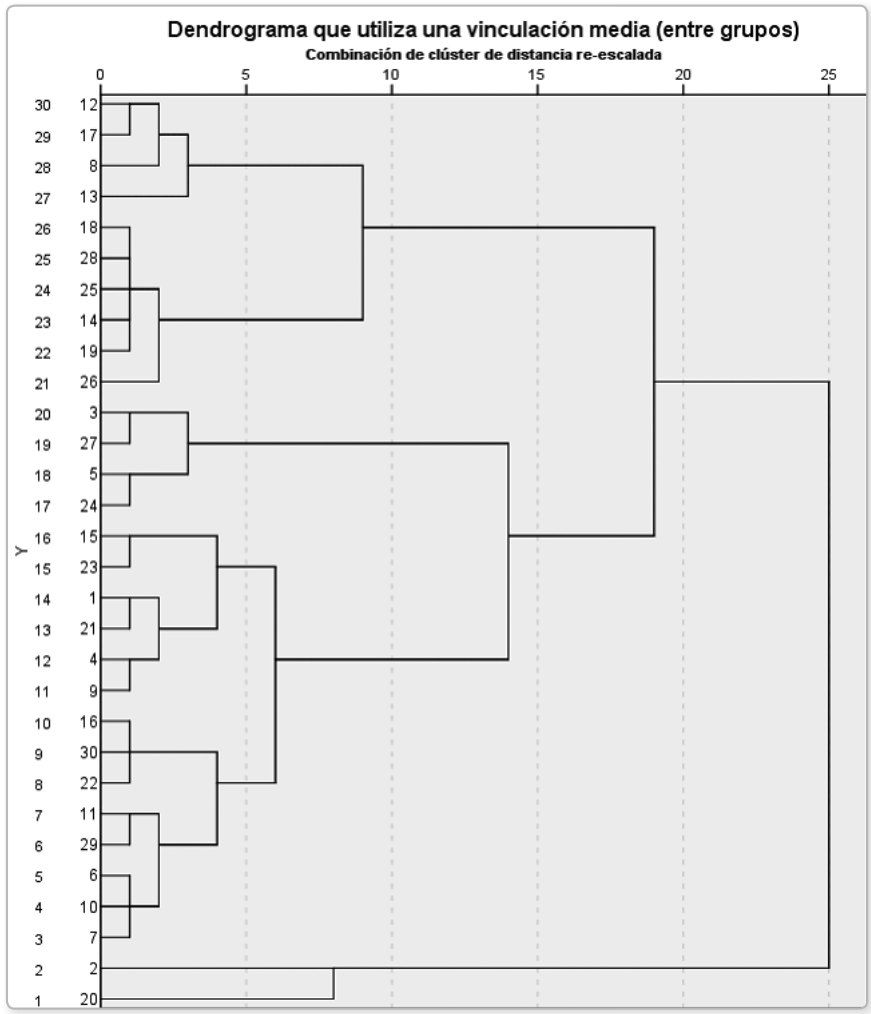
El cuadro siguiente, *Muestra el diagrama de témpanos,* resume el proceso de fusión de manera gráfica. En la cabecera de las columnas, se encuentran los números de los casos individuales (cada columna con un número que representa un caso) y en las filas el número de conglomerados formado en cada etapa (cada fila representa una etapa del proceso de fusión). Las etapas comienzan en la parte inferior del diagrama y van progresando hacia arriba.




Si hace clic en el cuadro Gráficos, y se obtiene el cuadro de diálogo siguiente:



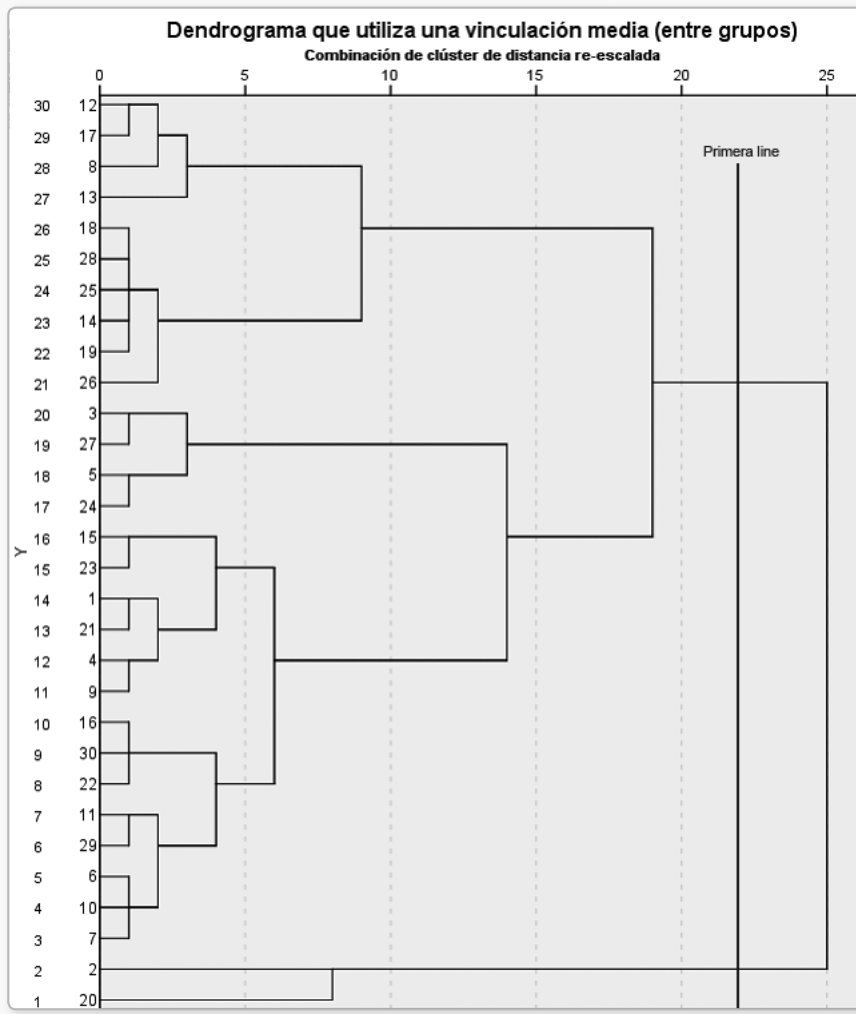
Luego, seleccione el dendrograma cuyo resultado es el gráfico que se muestra a continuación:




Este gráfico permite dividir los casos en dos, tres o más grupos, dependiendo de las necesidades y conocimientos del investigador. Para ello, trace una línea vertical sobre el gráfico del dendrograma.

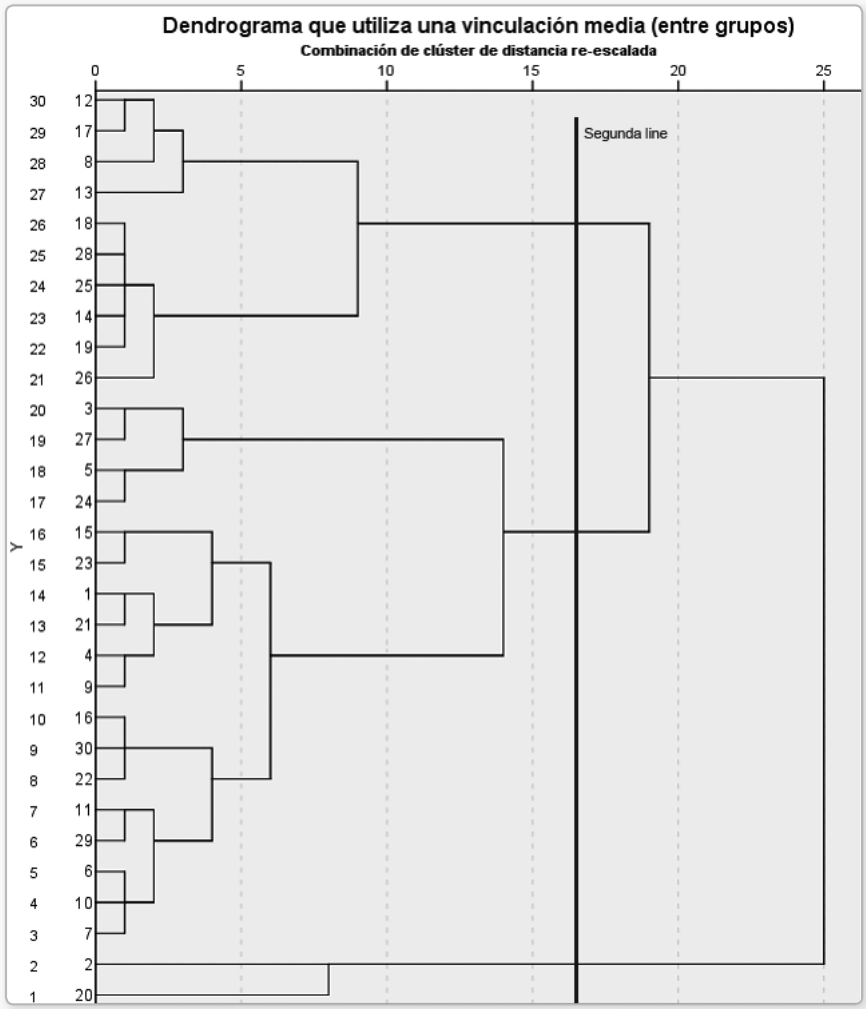
**Ejemplo 1:**

Si traza la primera línea, se obtienen dos grupos: el primer grupo formado por los valores 12, 17, 8, 13, 18…. 6, 10, 7 y el segundo grupo formado por casos 2 y 20. En consecuencia, la clasificación en dos grupos no es recomendable.



**Ejemplo 2:**

Si se traza la segunda línea, se obtienen tres grupos: el primer grupo formado por los valores 12, 17, 8, 13, 18… 14, 19, 26, el segundo grupo formado por casos 3, 27, 5… 6, 10, 7 y el tercer grupo por los casos 2 y 20. En consecuencia, la clasificación mejora enormemente.




Finalmente, se puede seguir clasificando en más grupos. Donde se produce la intersección de la línea trazada y las líneas de dendrograma se formarán los grupos. En consecuencia, si hay dos interceptos se formarán dos grupos, si hay tres interceptos se formarán tres grupos y así sucesivamente.

## C. ANÁLISIS DISCRIMINANTE

La discriminación o clasificación cuando se conocen los parámetros de las distribuciones, admiten una solución general. Un conjunto de elementos que pueden venir de dos o más poblaciones distintas. En cada elemento, se ha observado una variable aleatoria p−dimensional x, cuya distribución se conoce en las poblaciones consideradas. Se desea clasificar un nuevo elemento, con valores de las variables conocidas en una de las poblaciones.

El problema de discriminación aparece en muchas situaciones donde se necesita clasificar elementos con información incompleta. En otros casos, la información podría estar disponible, pero puede requerir destruir el elemento. Finalmente, en otros casos, la información puede ser muy costosa de adquirir. Las técnicas que se van a estudiar reciben también el nombre de «clasificación supervisada».

**Ejemplos:**

1. Los sistemas automáticos de concesión de créditos, implantados en instituciones financieras, tienen que utilizar variables (ingresos, antigüedad en el trabajo, patrimonio, etc.) para prever el comportamiento futuro.
2. En el reconocimiento de patrones para diseñar máquinas capaces de clasificar de manera automática, como reconocer voces y sonidos, clasificar billetes o monedas, reconocer caracteres escritos en una pantalla de ordenador o clasificar cartas según el distrito postal.
3. Asigne un texto escrito de procedencia desconocida a uno de varios autores, por las frecuencias de uso de palabras, asignar una partitura musical o un cuadro a un artista, una declaración de impuestos como potencialmente defraudadora o no, un paciente como enfermo de cáncer o no, un nuevo método de fabricación como eficaz o no.

Se recomienda antes de empezar a realizar un análisis discriminante, hacer un gráfico de dispersión, con la finalidad de tener una lectura inicial del comportamiento de los datos a analizar.

La figura en el espacio bivariante definido por las variables X1 y X2, las nubes de puntos correspondientes a dos grupos hipotéticos. Los dos grupos representados se diferencian entre sí en ambas variables, pero no por completo; pues, de hecho, se solapan en una pequeña región situada entre ambos. También está representada la función D, que es una combinación lineal de ambas variables. Sobre la función D, se representa la proyección de las dos nubes de puntos en forma de histograma, como si la función D cortase a las dos nubes de puntos en la dirección de su eje. Las dos líneas punteadas de cada uno de los histogramas representan la ubicación proyectada de los puntos medios de cada grupo (los centroides).

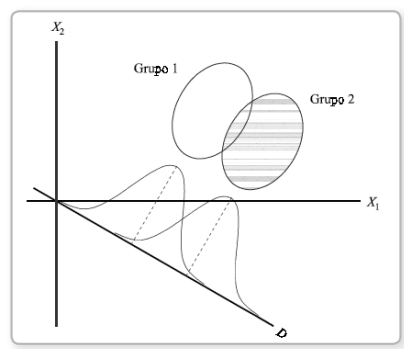



*Figura 7.1 Análisis discriminante clásico debido a Fisher*

Basada en la normalidad multivariante de las variables y que resulta óptimo bajo dicho supuesto. Si todas las variables son continuas, es frecuente que aunque los datos originales no sean normales, es posible transformar las variables para que lo sean y los métodos pueden aplicarse a las variables transformadas.

- **Poblaciones normales: Función lineal discriminante**

Sean $f_1$ y $f_2$ distribuciones normales con distintos vectores de medias, pero idéntica matriz de varianzas. Se desea clasificar un elemento genérico x que sí pertenece a la población i=1,2, y tiene función de densidad:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)' V^{-1}(x-\mu_i)\right\}$$

La partición óptima es clasificar en la población $P_2$ si:

$$\frac{f_2(x)\pi_2}{c(2|1)} > \frac{f_1(x)\pi_1}{c(1|2)}$$

Ambos términos son positivos, tomando logaritmos y sustituyendo $f_i(x)$.

$$-\frac{1}{2}(x-\mu_2)'V^{-1}(x-\mu_2) + \log\frac{\pi_2}{c(2|1)} > -\frac{1}{2}(x-\mu_1)'V^{-1}(x-\mu_1) + \log\frac{\pi_1}{c(1|2)}$$

Llamando $D_i^2$ a la distancia de Mahalanobis entre el punto observado (x) y la media de la población (i):

$$D_i^2 = (x-\mu_i)'V^{-1}(x-\mu_i)$$

$$D_1^2 - \log\frac{\pi_1}{c(1|2)} > D_2^2 - \log\frac{\pi_2}{c(2|1)}$$

Suponiendo que los costes son iguales y las probabilidades a priori: c(1/2) = c(2/1); $\pi_1 = \pi_2$; la regla anterior se reduce a:

Clasificar en 2 si $D_1^2 > D_2^2$

Es decir, clasificar la observación en la población de cuya media esté más próxima, midiendo la distancia con la medida de Mahalanobis. Se observa que si las variables x tuviesen $V = I\sigma^2$, la regla equivale a utilizar la distancia euclídea.

## Ejemplo 1:

Se desea clasificar un retrato entre dos posibles pintores. Para ello, se miden dos variables: la profundidad del trazo y la proporción que ocupa el retrato sobre la superficie del lienzo. Las medias de estas variables para el primer pintor (A) son (2 y 0.8), y para el segundo B, son (2.3 y 0.7) y las desviaciones típicas de estas variables son 0.5 y 0.1 y la correlación entre estas medidas es 0.5. La obra a clasificar tiene medidas de estas variables (2.1 y 0.75). Calcule las probabilidades de error.

Las distancias de Mahalanobis se darán calculando la covarianza como el producto de la correlación por las desviaciones típicas:

$$D_A^2 = (2.1-2,.75-.8)\begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1}\begin{pmatrix} 2.1-2 \\ 75-.8 \end{pmatrix} = 0.52$$

Para la segunda:

$$D_B^2 = (2.1-2.3,.75-.7)\begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1}\begin{pmatrix} 2.1-2.3 \\ .75-.7 \end{pmatrix} = 0,8133$$

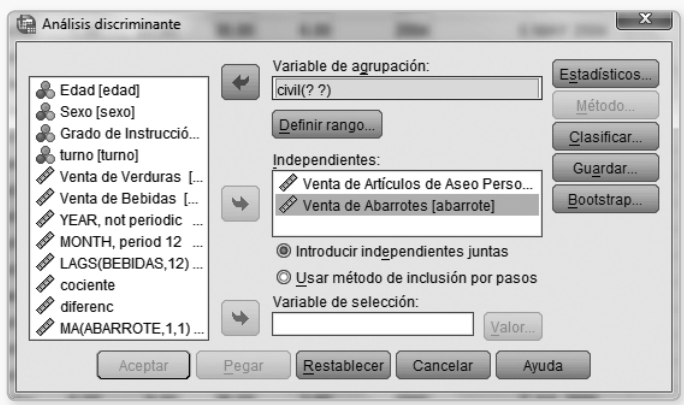
Por tanto, asigne la obra al primer pintor.

Para llevar a cabo un análisis discriminante en SPSS, seleccione la opción **Analizar** > **Clasificar** > **Análisis discriminante**, obteniéndose el cuadro de diálogo siguiente.

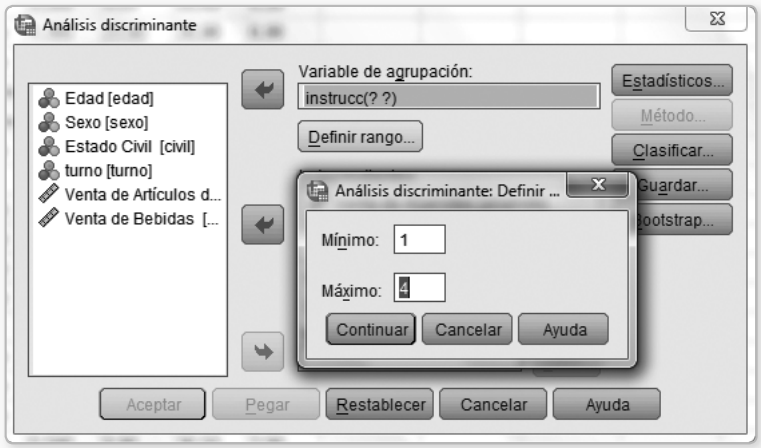



El cuadro de la izquierda muestra la lista de variables del archivo (excepto las que tienen formato cadena). Para obtener un análisis discriminante con las especificaciones que SPSS tiene establecidas, seleccione una variable categórica (nominal u ordinal) y traslade la variable de agrupación (Grouping Variable) al cuadro del lado derecho. Cabe aclarar que la **variable de agrupación** es aquella que define los grupos que se desea comparar; en este caso, será la variable Estado civil.

Luego, seleccione al menos una variable cuantitativa (intervalo o razón) y traslade a la lista independiente (independents). Las **variables independientes** son aquellas en las que se desea comparar los grupos, en este caso será la Venta de abarrotes y Venta de verduras (ver fig.).



Luego pulse el botón Definir rango... (Difine Range...) para acceder al cuadro de diálogo siguiente. En el ejemplo, la variable Estado civil tiene rango mínimo 1 y rango máximo 4.




Una vez ingresados los rangos, pulse el botón Continuar para volver al cuadro de diálogo principal. Luego de realizar estas selecciones, presione Aceptar (OK) y el visor presentará los siguientes resultados:

- El primer cuadro muestra un resumen con el total de los casos procesados, el número de casos válidos para el análisis y el número de casos excluidos.
- Dentro de los casos excluidos, se distinguen entre los que son excluidos porque su código en la variable de agrupación no esta dentro del rango seleccionado. Los que son excluidos por tener un valor perdido al menos en una variable discriminante y las que cumplen las dos condiciones anteriores.

**Analysis Case Processing Summary**

| Unweighted Cases | | N | Percent |
|---|---|---|---|
| Valid | | 30 | 100.0 |
| Excluded | Missing or out-of-range group codes | 0 | .0 |
| | At least one missing discriminating variable | 0 | .0 |
| | Both missing or out-of-range group codes and at least one missing discriminating variable | 0 | .0 |
| | Total | 0 | .0 |
| Total | | 30 | 100.0 |

El cuadro siguiente muestra un resumen del número de casos válidos en cada variable discriminante. La información de esta tabla posee un interés especial, pues un número desigual de casos en cada uno de los grupos puede afectar a la clasificación.

**Group Statistics**

| Estado Civil | | Valid N (listwise) Unweighted | Weighted |
|---|---|---|---|
| Soltero | Venta de Abarrotes | 5 | 5.000 |
| | Venta de Verduras | 5 | 5.000 |
| Casado | Venta de Abarrotes | 12 | 12.000 |
| | Venta de Verduras | 12 | 12.000 |
| Divorciado | Venta de Abarrotes | 4 | 4.000 |
| | Venta de Verduras | 4 | 4.000 |
| Viudo | Venta de Abarrotes | 9 | 9.000 |
| | Venta de Verduras | 9 | 9.000 |
| Total | Venta de Abarrotes | 30 | 30.000 |
| | Venta de Verduras | 30 | 30.000 |

El cuadro siguiente contiene los autovalores y algunas estadísticas descriptivas multivariantes. En la tabla se muestra una fila numerada por cada función discriminante. En este ejemplo hay dos funciones: la primera función explica el 89.8% de las diferencias existentes entre los sujetos de los grupos, y la segunda función explica 10.2% de las diferencias existentes entre los sujetos de los grupos.

**Eigenvalues**

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|---|---|---|---|---|
| 1 | .182[a] | 89.8 | 89.8 | .392 |
| 2 | .021[a] | 10.2 | 100.0 | .142 |

a. First 2 canonical discriminant functions were used in the analysis.

*Figura 7.2 [% de variancia (% of Variance) % acumulado (Cumulative %)]*

El autovalor (Eigenvalue) permite comparar cómo se distribuye la dispersión intergrupos cuando existe más de una función. Un autovalor tiene un valor mínimo de cero, pero no tiene un máximo, lo cual lo hace difícilmente interpretable por sí solo. Por esta razón, se utiliza la estadística lambda de Wilks que se encuentra estrechamente relacionada con los autovalores.

La **correlación canónica** (canonical Correlation) es la correlación entre la combinación lineal de las variables independientes (función discriminante) y una combinación lineal de variables indicador que recogen la pertenencia de los sujetos a los grupos. Una correlación canónica alta indica que las variables discriminantes permiten diferenciar entre los grupos.

En este caso, el autovalor es próximo a cero y la correlación canónica es moderada, por lo que debe suponerse que las variables discriminantes utilizadas (Venta de abarrotes y Venta de verduras) no permiten distinguir claramente entre los grupos en ambas funciones.

El cuadro siguiente muestra el estadístico *Lambda de Wilks,* que expresa la proporción de variabilidad total no debida a las diferencias entre los grupos. Permite contrastar la hipótesis nula de que las medias multivariantes de los grupos (los centroides) son iguales.

**Wilks' Lambda**

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 through 2 | .829 | 4.872 | 6 | .560 |
| 2 | .980 | .533 | 2 | .766 |

Por tanto, los valores de lambda de Wilks cercanos a ***1*** indican un gran parecido entre grupo, mientras que valores próximo a ***cero*** indican una gran diferencia entre ellos. En el ejemplo, los valores son cercanos a 1, lo cual indica un gran parecido entre grupos o la existencia de bastante solapamiento entre los grupos. Sin embargo, el valor transformado de *lambda (Chi-Square)* es de 4.872 con 6 grados de libertad y un nivel crítico (Sig) de 0.560, por lo que se puede aceptar la hipótesis nula acerca de que los grupos comparados tienen promedios iguales en las dos variables discriminantes en la primera función.

Para la segunda función el *Chi-Square* = 0.533 con 2 grados de libertad y un nivel crítico (Sig) de 0.766, por lo que se puede aceptar la hipótesis nula de que los grupos comparados tienen promedios iguales en las dos variables discriminantes. En consecuencia, los grupos son parecidos y no podría realizar un análisis discriminante.

El cuadro siguiente muestra la tabla de *coeficientes estandarizados,* que contiene una versión estandarizada de los coeficientes de la función canónica discriminante.

**Standardized Canonical Discriminant Function Coefficients**

| | Function | |
|---|---|---|
| | 1 | 2 |
| Venta de Abarrotes | 1.003 | .163 |
| Venta de Verduras | .016 | 1.016 |

En la función 1, la venta de abarrotes tiene mayor importancia que las ventas de verduras, y en la función 2, la venta de verduras tiene mayor importancia que las ventas de abarrotes.

El cuadro siguiente muestra la *matriz de estructura,* que contiene las correlaciones entre las variables discriminantes y la función discriminante estandarizada.

**Structure Matrix**

| | Function | |
|---|---|---|
| | 1 | 2 |
| Venta de Abarrotes | 1.000* | -.016 |
| Venta de Verduras | -.161 | .987* |

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within functio

*. Largest absolute correlation between each variable and any discriminant function

Se puede apreciar que la venta de abarrotes correlaciona alto con la función discriminante uno, y que la venta de verduras correlaciona alto con la función discriminante dos.

El cuadro siguiente muestra la ubicación de los *centroides* en la función discriminante. Esta tabla es de gran utilidad para interpretar la función discriminante.

**Functions at Group Centroids**

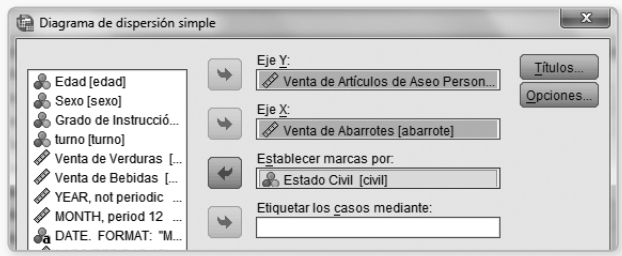
| | Function | |
|---|---|---|
| Estado Civil | 1 | 2 |
| Soltero | .886 | -1.43E-02 |
| Casado | -.157 | .153 |
| Divorciado | -.176 | -7.67E-02 |
| Viudo | -.204 | -.162 |

Unstandardized canonical discriminant functions evaluated at group means

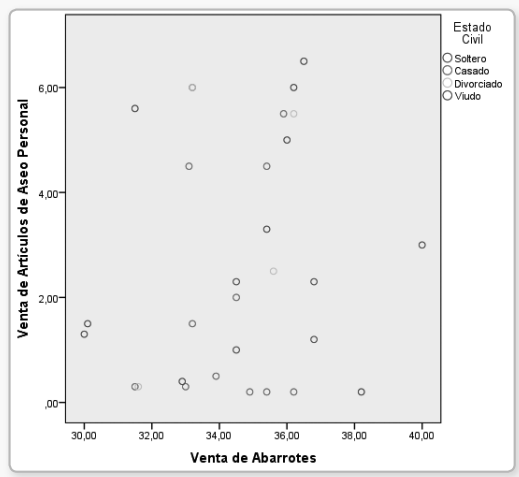
Se observa que el grupo de solteros se encuentran localizados, en promedio, en las puntuaciones (ventas) positivas de la primera función, mientras que los casados, divorciados y viudos se encuentran en las puntuaciones (ventas) negativas, respecto de la primera función.

### Diagrama de dispersión

Es importante realizar un diagrama de dispersión para poder tener una primera lectura de los grupos que se han seleccionado. Es necesario recordar que los grupos se seleccionan en análisis discriminante.




Se obtiene el siguiente gráfico:




Se observa que no se pueden distinguir claramente los grupos; en consecuencia, no se podrá realizar un análisis discriminante. Aquí reside la importancia de realizar un gráfico de dispersión.

## 7.2. REDUCCIÓN DE DATOS

Técnica estadística que permite la reducción de datos, que sirve para encontrar grupos homogéneos. También permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables.

SPSS dispone de los siguientes procedimientos: *Análisis factorial, Análisis de correspondencias y Escalamiento óptimo,* que se definen a continuación:




### 7.2.1. ANÁLISIS FACTORIAL

Intenta identificar variables subyacentes o factores que expliquen la configuración de las correlaciones, dentro de un conjunto de variables observadas. Se suele utilizar en la reducción de los datos, para identificar un pequeño número de factores que explique la mayoría de la varianza observada en un número mayor de variables manifiestas.

También puede utilizarse para generar hipótesis relacionadas con los mecanismos causales, o para inspeccionar las variables de análisis subsiguientes (por ejemplo, para identificar la colinealidad, antes de realizar un análisis de regresión lineal). Este procedimiento dispone de siete métodos de extracción factorial; 5 métodos de rotación, entre ellos, oblimin directo y promax para rotaciones no ortogonales; y 3 métodos para calcular las puntuaciones factoriales que pueden guardarse como variables de análisis adicionales.

### 7.2.2. ANÁLISIS DE CORRESPONDENCIAS

Describe las relaciones existentes entre dos variables nominales, recogidas en una tabla de correspondencias, sobre un espacio de pocas dimensiones; mientras que, al mismo tiempo, se describen las relaciones entre las categorías de cada variable.

Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las categorías, con las categorías similares representadas próximas unas a otras. La proyección de los puntos de una variable sobre el vector, desde el origen hasta un punto de categoría de la otra variable, describe la relación entre ambas variables. El análisis de las tablas de contingencia, a menudo, incluye examinar los perfiles de fila y de columna, así como contrastar la independencia por medio del estadístico de chi-cuadrado. Sin embargo, el número de perfiles puede ser muy grande y la prueba de chi-cuadrado no revelará la estructura de la dependencia.

El procedimiento Tablas de contingencia ofrece varias medidas y pruebas de asociación, pero no puede representar gráficamente alguna relación entre las variables. El **análisis factorial** es una técnica típica para describir las relaciones existentes entre variables, en un espacio de pocas dimensiones; sin embargo, requiere datos de intervalo y el número de observaciones debe ser cinco veces el número de variables.

Por su parte, el análisis de correspondencias asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además, el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas.

### 7.2.3. ESCALAMIENTO ÓPTIMO

La combinación del nivel de medida y el número de conjunto seleccionado determina este procedimiento mediante los mínimos cuadrados alternantes que realiza.

a. **Nivel de medida**. Permite especificar el nivel correspondiente a las variables utilizadas en el análisis:
   - **Todas las variables son nominales múltiples:** Todas las variables tienen cuantificaciones de categorías que pueden diferir para cada dimensión.
   - **Alguna variable no es nominal múltiple:** Una o más variables se escalan a un nivel diferente del nominal múltiple. Otros niveles de escala posibles son nominal simple, ordinal y numérica discreta.

b. **Número de conjuntos de variables.** Permite especificar cuántos grupos de variables se van a comparar con otros grupos de variables:
   - **Un conjunto**: Los datos contienen un grupo de variables.
   - **Múltiples conjuntos:** Los datos contienen más de un grupo de variables. Si se selecciona esta opción, se elegirá Correlación canónica no-lineal (OVERALS).

c. **Análisis seleccionado.** Las opciones para Nivel de medida y Número de conjuntos de variables deciden el análisis según las combinaciones.

d. **Análisis de homogeneidad (HOMALS).** Seleccionando Todas las variables son nominales múltiples y Un conjunto:
   - **Análisis de componentes principales categóricos (CATPCA):** Seleccionando Alguna variable no es nominal múltiple y Un conjunto.
   - **Análisis de correlación canónica no lineal (OVERALS):** Seleccionando Múltiples conjuntos.

### EJERCICIO PRÁCTICO N.° 1

#### A. Componentes principales

Si es posible describir con precisión los valores de p variables por un pequeño subconjunto r < p de ellas, se habrá reducido la dimensión a costa de una pequeña pérdida de información. Dada n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Su utilidad es doble:

- Permite representar observaciones de un espacio general p-dimensional en un espacio de dimensión pequeña. Es el primer paso para identificar posibles variables latentes o no observadas que están generando la variabilidad de los datos.
- Permite transformar las variables originales, en general correlacionadas, en nuevas variables incorrelacionadas, facilitando la interpretación de los datos.

#### a. Selección del número de componentes

Entre las principales formas de seleccionar, se tienen las siguientes tres componentes:

- Realice un gráfico de $\lambda_i$ frente a i. Seleccione componentes hasta que los restantes tengan aproximadamente el mismo valor de $\lambda_i$. El objetivo es buscar un codo en el gráfico; es decir, un punto a partir del cual los valores propios son aproximadamente iguales.

- Seleccione componentes hasta cubrir una proporción determinada de varianza, como 80% o 90%. Debe aplicarse con cierto cuidado; por ejemplo, es posible que un único componente recoja el 90% de la variabilidad y, sin embargo, pueden existir otros componentes que sean muy adecuados para explicar la forma de las variables.
- Deseche los componentes asociados a valores propios, inferiores a una cota que suele fijarse como la varianza media $\sum\lambda_i/p$. Para la matriz de correlación, el valor medio es 1. Seleccione los valores propios mayores a 1 (regla arbitraria: una variable independiente del resto suele llevarse un componente principal y puede tener un valor propio mayor que la unidad; sin embargo, si está incorrelada con el resto, puede ser una variable poco relevante para el análisis).

### B. Análisis factorial

Técnica estadística multivariante, cuyo principal propósito es sintetizar las interrelaciones observadas entre un conjunto de variables, en forma concisa y segura, como una ayuda a la construcción de nuevos conceptos y teorías. Para ello, se utiliza un conjunto de variables aleatorias inobservables llamadas «factores comunes», de forma que todas las covarianzas o correlaciones son explicadas por dichos factores y cualquier porción de la varianza inexplicada, por los factores comunes, se asigna a términos de error residuales llamados «factores únicos o específicos». El análisis factorial puede ser *exploratorio o confirmatorio*: el análisis exploratorio se caracteriza porque no se conocen a priori el número de factores y es, en la aplicación empírica, donde se determina este número. Por el contrario, en el análisis de tipo confirmatorio, los factores están fijados a priori, utilizándose contrastes de hipótesis para su corroboración.

A continuación, se va a realizar un análisis factorial, utilizando el método de análisis de componentes principales. El presente ejercicio consiste en resolver un problema por medio de la técnica estadística de análisis de componentes principales. Los datos representan una muestra de 41 ciudades del Perú en las que se midieron diferentes variables relacionadas con la contaminación atmosférica.

| | $SO_2$ | Temperatura | Empresas | Población | Viento | Precipitación | Días |
|---|---|---|---|---|---|---|---|
| Iquitos | 10,00 | 70,30 | 213,00 | 582,00 | 6,00 | 7,05 | 36,00 |
| Sicuani | 13,00 | 61,00 | 91,00 | 132,00 | 8,20 | 48,52 | 100,00 |
| Huánuco | 12,00 | 56,70 | 453,00 | 716,00 | 8,70 | 20,66 | 67,00 |
| Trujillo | 17,00 | 51,90 | 454,00 | 515,00 | 9,00 | 12,95 | 86,00 |
| Chiclayo | 56,00 | 49,10 | 412,00 | 158,00 | 9,00 | 43,37 | 127,00 |
| Paita | 36,00 | 54,00 | 80,00 | 80,00 | 9,00 | 40,25 | 114,00 |
| Pucallpa | 29,00 | 57,30 | 434,00 | 757,00 | 9,30 | 38,89 | 111,00 |
| Cusco | 14,00 | 68,40 | 136,00 | 529,00 | 8,80 | 54,47 | 116,00 |
| Huancayo | 10,00 | 75,50 | 207,00 | 335,00 | 9,00 | 59,80 | 128,00 |
| Chimbote | 24,00 | 61,50 | 368,00 | 497,00 | 9,10 | 48,34 | 115,00 |
| Lima | 110,00 | 50,60 | 3344,00 | 3369,00 | 10,40 | 34,44 | 122,00 |
| Tacna | 28,00 | 52,30 | 361,00 | 746,00 | 9,70 | 38,74 | 121,00 |
| Sullana | 17,00 | 49,00 | 104,00 | 201,00 | 11,20 | 30,85 | 103,00 |
| Juliaca | 8,00 | 56,60 | 125,00 | 277,00 | 12,70 | 30,58 | 82,00 |
| Ica | 30,00 | 55,60 | 291,00 | 593,00 | 8,30 | 43,11 | 123,00 |
| Cajamarca | 9,00 | 68,30 | 204,00 | 361,00 | 8,40 | 56,77 | 113,00 |
| Chincha | 47,00 | 55,00 | 625,00 | 905,00 | 9,60 | 41,31 | 111,00 |
| Arequipa | 35,00 | 49,90 | 1064,00 | 1513,00 | 10,10 | 30,96 | 129,00 |
| Tumbes | 29,00 | 43,50 | 699,00 | 744,00 | 10,60 | 25,94 | 137,00 |
| Ayacucho | 14,00 | 54,50 | 381,00 | 507,00 | 10,00 | 37,00 | 99,00 |
| Puno | 56,00 | 55,90 | 775,00 | 622,00 | 9,50 | 35,89 | 105,00 |
| Tarapoto | 14,00 | 51,50 | 181,00 | 347,00 | 10,90 | 30,18 | 98,00 |
| Huaraz | 11,00 | 56,80 | 46,00 | 244,00 | 8,90 | 7,77 | 58,00 |
| Talara | 46,00 | 47,60 | 44,00 | 116,00 | 8,80 | 33,36 | 135,00 |
| Huaral | 11,00 | 47,10 | 391,00 | 463,00 | 12,40 | 36,11 | 166,00 |
| Jaén | 23,00 | 54,00 | 462,00 | 453,00 | 7,10 | 39,04 | 132,00 |
| Pasco | 65,00 | 49,70 | 1007,00 | 751,00 | 10,90 | 34,99 | 155,00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Chulucanas | 26,00 | 51,50 | 266,00 | 540,00 | 8,60 | 37,01 | 134,00 |
| Piura | 69,00 | 54,60 | 1692,00 | 1950,00 | 9,60 | 39,93 | 115,00 |
| Catacaos | 61,00 | 50,40 | 347,00 | 520,00 | 9,40 | 36,22 | 147,00 |
| Moyobamba | 94,00 | 50,00 | 343,00 | 179,00 | 10,60 | 42,75 | 125,00 |
| Lambayeque | 10,00 | 61,60 | 337,00 | 624,00 | 9,20 | 49,10 | 105,00 |
| Yurimaguas | 18,00 | 59,40 | 275,00 | 448,00 | 7,90 | 46,00 | 119,00 |
| Barranca | 9,00 | 66,20 | 641,00 | 844,00 | 10,90 | 35,94 | 78,00 |
| Callao | 10,00 | 68,90 | 721,00 | 1233,00 | 10,80 | 48,19 | 103,00 |
| Tarma | 28,00 | 51,00 | 137,00 | 176,00 | 8,70 | 15,17 | 89,00 |
| Ilo | 31,00 | 59,30 | 96,00 | 308,00 | 10,60 | 44,68 | 116,00 |
| Huacho | 26,00 | 57,80 | 197,00 | 299,00 | 7,60 | 42,59 | 115,00 |
| Abancay | 29,00 | 51,10 | 379,00 | 531,00 | 9,40 | 38,79 | 164,00 |
| Sihuas | 31,00 | 55,20 | 35,00 | 71,00 | 6,50 | 40,75 | 148,00 |
| Tingo María | 16,00 | 45,70 | 569,00 | 717,00 | 11,80 | 29,07 | 123,00 |

Las variables son las siguientes:

- Contenido en $SO_2$ (óxido de azufre, llamado «dióxido de azufre»)
- Temperatura anual en grados F
- Número de empresas mayores de 20 trabajadores
- Población (en miles de habitantes)
- Velocidad media del viento
- Precipitación anual media
- Días lluviosos al año

Interesa investigar la relación entre la concentración en $SO_2$ y el resto de variables (para eliminar relaciones entre las variables, se emplea un análisis de componentes principales).

a. Realice un análisis de los componentes principales, sobre todas las variables, salvo $SO_2$.
b. Haga gráficas descriptivas exploratorias y observe los datos anómalos (outliers).
c. Ejecute la prueba de esfericidad de Barlett.
d. Efectúe gráficos de componentes.

## Solución:

Se realiza un análisis de componentes principales sobre todas las variables, excepto $SO_2$.

### A. Análisis de componentes principales (con SPSS)

El objetivo del análisis de componentes principales es identificar, a partir de un conjunto de *p* variables, otro conjunto de *k* ($k < p$) variables no directamente observables, denominadas «factores», tal que:

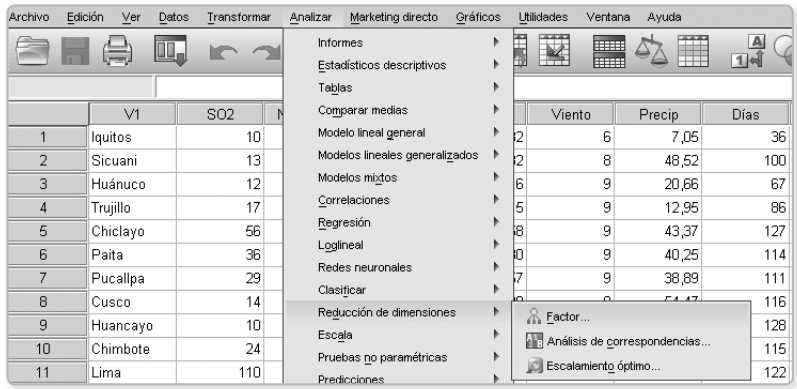
- *k* sea un número pequeño.
- Se pierda la menor cantidad posible de información.
- La solución obtenida sea interpretable.

Realice los siguientes pasos en el análisis de componentes principales:

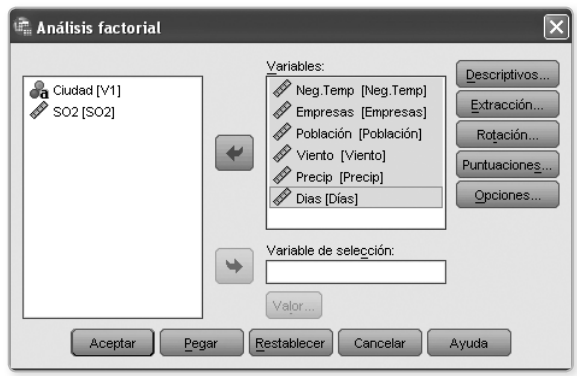
- Evaluación de lo apropiado de realizar el análisis.
- Extracción de los factores.
- Cálculo de las puntuaciones factoriales para cada caso.

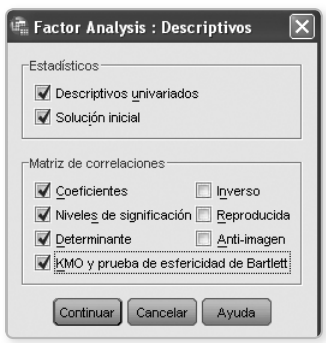
Dentro del menú Analizar se encuentra el procedimiento **Reducción de datos**. El aspecto de dicho menú es el siguiente:

Analizar > Reducción de datos (Data Reduction) > Factor...



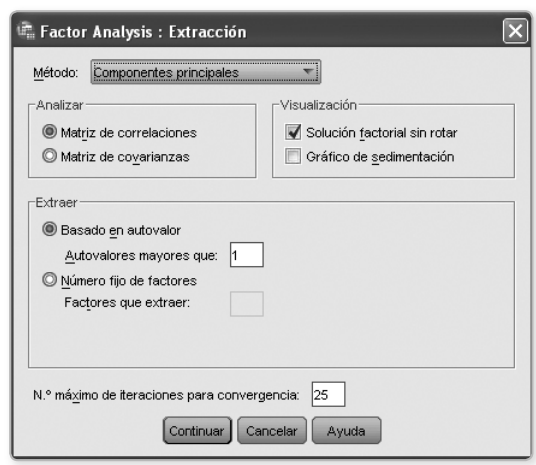

Se muestra un cuadro de diálogo: al cuadro de variables se pasan las seis variables, de las que se está interesado investigar la relación entre la concentración en $SO_2$.



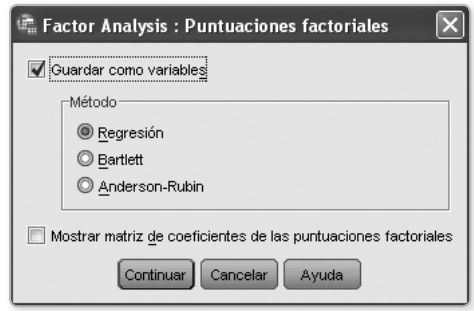




A continuación, se ingresa a descriptivos y se marcan las siguientes estadísticas:

Luego haga clic en Extracción:



Finalmente:




Se obtienen los siguientes resultados:

**1.** La tabla representa las medias y desviaciones estándares de cada variable:

**Estadísticos descriptivos**

| | Media | Desviación típica | N del análisis |
|---|---|---|---|
| Neg.Temp | 55,763 | 7,2277 | 41 |
| Empresas | 463,10 | 563,474 | 41 |
| Población | 608,61 | 579,113 | 41 |
| Viento | 9,44 | 1,429 | 41 |
| Precip | 36,7690 | 11,77155 | 41 |
| Días | 113,90 | 26,506 | 41 |

**2.** Se calcula la matriz de correlaciones con la significación de cada componente. Se obtienen los componentes principales a partir de la matriz de correlaciones, para emplear las mismas escalas en todas las variables.

**Matriz de correlaciones[a]**

| | | Neg.Temp | Empresas | Población | Viento | Precip | Días |
|---|---|---|---|---|---|---|---|
| Correlación | Neg.Temp | 1,000 | -,190 | -,063 | -,350 | ,386 | -,430 |
| | Empresas | -,190 | 1,000 | ,955 | ,238 | -,032 | ,132 |
| | Población | -,063 | ,955 | 1,000 | ,213 | -,026 | ,042 |
| | Viento | -,350 | ,238 | ,213 | 1,000 | -,013 | ,164 |
| | Precip | ,386 | -,032 | -,026 | -,013 | 1,000 | ,496 |
| | Días | -,430 | ,132 | ,042 | ,164 | ,496 | 1,000 |
| Sig. (Unilateral) | Neg.Temp | | ,117 | ,349 | ,012 | ,006 | ,002 |
| | Empresas | ,117 | | ,000 | ,067 | ,420 | ,206 |
| | Población | ,349 | ,000 | | ,091 | ,436 | ,397 |
| | Viento | ,012 | ,067 | ,091 | | ,468 | ,153 |
| | Precip | ,006 | ,420 | ,436 | ,468 | | ,000 |
| | Días | ,002 | ,206 | ,397 | ,153 | ,000 | |

a. Determinante = ,014

Para que se pueda realizar el análisis de componentes principales, es necesario que las variables presenten factores comunes; es decir, que estén muy correlacionadas entre sí. Los coeficientes de la matriz de correlaciones deben ser grandes en valor absoluto.

3. **Test de esfericidad de Barlett.** Para comprobar que las correlaciones entre las variables son distintas de cero de modo significativo, se comprueba si el determinante de la matriz es distinto de uno; es decir, si la matriz de correlaciones es distinta de la matriz identidad.

   Si las variables están correlacionadas, hay muchos valores altos en valor absoluto fuera de la diagonal principal de la matriz de correlaciones; además, el determinante es menor que 1 (el máximo valor del determinante es 1 si las variables están incorrelacionadas).

   El test de Barlett realiza el siguiente contraste:

   $$H_0: |R| = 1$$
   $$H_1: |R| \neq 1$$

   El determinante de la matriz da una idea de la correlación generalizada entre todas las variables. Se basa el test en la distribución chi-cuadrado, donde valores altos llevan a rechazar $H_0$; así, la prueba de esfericidad de Bartlett contrasta si la matriz de correlaciones es una matriz identidad, que indicaría que el modelo factorial es inadecuado. Por otro lado, la medida de la adecuación muestral de Kaiser-Meyer-Olkin contrasta si las correlaciones parciales entre las variables son pequeñas.

**KMO y prueba de Bartlett**

| | | |
|---|---|---|
| Medida de adecuación muestral de Kaiser-Meyer-Olkin. | | ,365 |
| Prueba de esfericidad de Bartlett | Chi-cuadrado aproximado | 159,231 |
| | gl | 15 |
| | Sig. | ,000 |

4. **Comunalidad.** Está asociada con la variable j-ésima. Es la proporción de variabilidad de dicha variable explicada por los *k* factores considerados.

   Equivale a la suma de la fila j-ésima de la matriz factorial. Sería igual a 0 si los factores comunes no explicasen algo la variabilidad de una variable, y sería igual a 1 si quedase totalmente explicada.

**Comunalidades**

| | Inicial | Extracción |
|---|---|---|
| Neg.Temp | 1,000 | ,892 |
| Empresas | 1,000 | ,968 |
| Población | 1,000 | ,979 |
| Viento | 1,000 | ,424 |
| Precip | 1,000 | ,941 |
| Días | 1,000 | ,888 |

Método de extracción: Análisis de Componentes principales.

   La variable Viento aporta poca variabilidad al modelo, que es de 0.424 y no es muy significativa.

5. **Varianza total explicada.** Se utiliza para determinar cuántos factores deben retenerse. Los primeros tres componentes tienen todos varianzas (autovalores) mayores que 1, y entre los tres recogen el 84.846% de la varianza de las variables originales.

**Varianza total explicada**

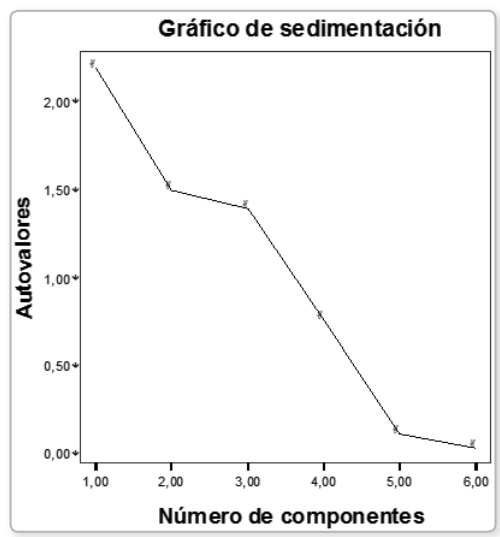
| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 2,196 | 36,603 | 36,603 | 2,196 | 36,603 | 36,603 |
| 2 | 1,500 | 24,999 | 61,602 | 1,500 | 24,999 | 61,602 |
| 3 | 1,395 | 23,244 | 84,846 | 1,395 | 23,244 | 84,846 |
| 4 | ,760 | 12,670 | 97,516 | | | |
| 5 | ,115 | 1,910 | 99,426 | | | |
| 6 | ,034 | ,574 | 100,000 | | | |

Método de extracción: Análisis de Componentes principales.

Puede observar que la primera componente aporta un 36.603% de variabilidad, mientras que la segunda componente aporta 24.999% de variabilidad; sumando la variabilidad de la primera y segunda componentes se obtiene 61.602% de variabilidad, pero si suma la variabilidad de la tercera componente (23.244%) se obtiene 84.846% de variabilidad.

6. **Gráfico de sedimentación**. Gráfico de la varianza asociada a cada factor. Se utiliza para determinar cuántos factores deben retenerse. Típicamente, el gráfico muestra la clara ruptura entre la pronunciada pendiente de los factores más importantes y el descenso gradual de los restantes (los sedimentos).

Otra opción es usar el criterio de Kaiser, que consiste en conservar aquellos factores cuyo autovalor asociado sea mayor que 1.



7. **Matriz de componentes.** El primer componente se podría etiquetar como calidad de vida con valores negativos altos en empresas y población, indicando un entorno relativamente pobre. El segundo componente se puede etiquetar como tiempo húmedo, y tiene pesos altos en las variables precipitaciones y días. El tercer componente se podría etiquetar como tipo de clima, y está relacionado con la temperatura y la cantidad de lluvia. Aunque no se encontrasen etiquetas claras para los componentes, siempre es interesante calcular componentes principales para descubrir si los datos se encuentran en una dimensión menor; de hecho, los tres primeros componentes producen un mapa de los datos, donde las distancias entre los puntos es bastante semejante a la observada en los mismos, respecto a las variables originales.

**Matriz de componentes[a]**

| | Componente | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Neg.Temp | -,489 | -,156 | ,793 |
| Empresas | ,906 | -,206 | ,322 |
| Población | ,856 | -,272 | ,414 |
| Viento | ,524 | ,160 | -,351 |
| Precip | -,060 | ,763 | ,596 |
| Días | ,353 | ,867 | -,110 |

Método de extracción: Análisis de componentes principales.

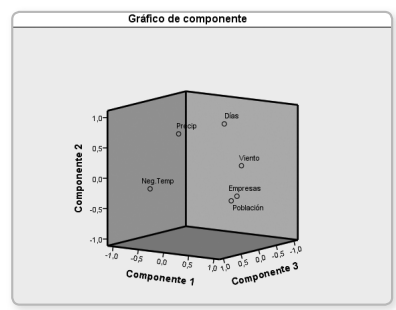
a. 3 componentes extraídos

8. **Matriz de correlaciones estimada a partir de la solución factorial.** También se muestran las correlaciones residuales (la diferencia entre la correlación observada y la reproducida).
   - Correlaciones reproducidas:

| | | Neg. Temp | Empresas | Población | Viento | Precio | Días |
|---|---|---|---|---|---|---|---|
| Correlación reproducida | Neg.temp | ,892(b) | -,155 | -,048 | -,560 | ,383 | -,395 |
| | Empresas | -,155 | ,968(b) | ,965 | ,329 | -,020 | ,106 |
| | Población | -,048 | ,965 | ,979(b) | ,260 | -,013 | ,020 |
| | Viento | -,560 | ,329 | ,260 | ,424(b) | -,119 | ,362 |
| | Precip | ,383 | -,020 | -,013 | -,119 | ,941(b) | ,574 |
| | Días | -,395 | ,106 | ,020 | ,362 | ,574 | ,888(b) |
| Residual(a) | Neg.temp | | -,035 | -,015 | ,210 | ,003 | -,035 |
| | Empresas | -,035 | | -,010 | -,091 | -,013 | ,026 |
| | Población | -,015 | -,010 | | -,047 | -,013 | ,022 |
| | Viento | ,210 | -,091 | -,047 | | ,106 | -,198 |
| | Precip | ,003 | -,013 | -,013 | ,106 | | -,078 |
| | Días | -,035 | ,026 | ,022 | -,198 | -,078 | |

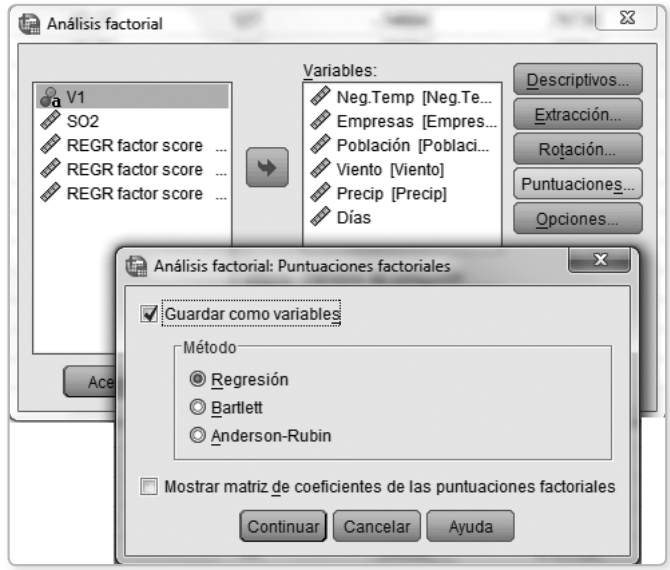
*Figura 7.3 Método de extracción: Análisis de componentes principales: (a) Los residuos se calculan entre las correlaciones observadas y reproducidas; hay 5 (33,0%) residuales no redundantes con valores absolutos mayores que 0,05. (b) Comunalidades reproducidas.*

9. **Gráfico de saturación**




En el gráfico, se observa que existe una correlación positiva entre la variable Empresa y la variable Población.

**10. Guardar los factores.** Si desea guardar los factores, ingrese al cuadro de diálogo de puntaciones y marque Guardar puntuaciones.




Luego, se obtienen las puntuaciones de cada factor en la Vista de datos, como se muestra en el gráfico siguiente:

| V1 | SO2 | Neg.Temp | Empresas | Población | Viento | Precip | Días | FAC1_1 | FAC2_1 | FAC3_1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Iquitos | 10 | 70,3 | 213 | 582 | 6 | 7,05 | 36 | -1,62635 | -3,38013 | ,78750 |
| Sicuani | 13 | 61,0 | 91 | 132 | 8 | 48,52 | 100 | -1,07415 | ,27621 | ,70232 |
| Huánuco | 12 | 56,7 | 453 | 716 | 9 | 20,66 | 67 | -,33464 | -1,81887 | -,18956 |
| Trujillo | 17 | 51,9 | 454 | 515 | 9 | 12,95 | 86 | -,13826 | -1,58331 | -1,05904 |
| Chiclayo | 56 | 49,1 | 412 | 158 | 9 | 43,37 | 127 | -,14604 | ,78738 | -,49732 |
| Paita | 36 | 54,0 | 80 | 80 | 9 | 40,25 | 114 | -,66394 | ,40384 | -,36242 |
| Pucallpa | 29 | 57,3 | 434 | 757 | 9 | 38,89 | 111 | -,01528 | -,04401 | ,29597 |
| Cusco | 14 | 68,4 | 136 | 529 | 9 | 54,47 | 116 | -,81837 | ,68482 | 1,56915 |
| Huancayo | 10 | 75,5 | 207 | 335 | 9 | 59,80 | 128 | -1,02186 | 1,13288 | 2,18012 |
| Chimbote | 24 | 61,5 | 368 | 497 | 9 | 48,34 | 115 | -,39924 | ,47360 | ,83255 |
| Lima | 110 | 50,6 | 3344 | 3369 | 10 | 34,44 | 122 | 4,34156 | -1,34554 | 1,91222 |
| Tacna | 28 | 52,3 | 361 | 746 | 10 | 38,74 | 121 | ,20550 | ,29073 | -,23872 |
| Sullana | 17 | 49,0 | 104 | 201 | 11 | 30,85 | 103 | -,08797 | -,04936 | -1,38012 |
| Juliaca | 8 | 56,6 | 125 | 277 | 13 | 30,58 | 82 | -,13118 | -,54525 | -,94614 |
| Ica | 30 | 55,6 | 291 | 593 | 8 | 43,11 | 123 | -,28243 | ,43595 | ,31320 |
| Cajamarca | 9 | 68,3 | 204 | 361 | 8 | 56,77 | 113 | -,96901 | ,72645 | 1,66597 |
| Chincha | 47 | 55,0 | 625 | 905 | 10 | 41,31 | 111 | ,33951 | ,02319 | ,30411 |

**C. Análisis de correspondencias**

Es una técnica descriptiva para representar tablas de contingencia, que son tablas donde se recogen las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos. La información de partida ahora es una matriz de dimensiones IxJ, que representa las frecuencias absolutas observadas de dos variables cualitativas en n elementos. La primera variable se representa por flas y se supone que toma **I** valores posibles, en tanto que la segunda se representa por columnas, tomando **J** valores posibles.

Interpretación: La primera, como una manera de representar las variables en un espacio de dimensión menor de forma análoga a componentes principales, pero definiendo la distancia entre los puntos de manera coherente, con la interpretación de los datos y, en lugar de utilizar la distancia euclídea, se usa la distancia ji-cuadrado. La segunda interpretación está más próxima al escalado multidimensional, se trata de un procedimiento objetivo de asignar valores numéricos a variables cualitativas.

## Ejemplo:

Se presenta la clasificación de n = 5387 escolares por el color de sus ojos, que tiene cuatro categorías posibles (I = 4), y el color de su cabello, que tiene cinco categorías posibles (J = 5).

| Color de ojos | Color del pelo | | | | | |
|---|---|---|---|---|---|---|
| | Rubio | Pelirrojo | Castaño | Oscuro | Negro | **Total** |
| Claros | 688 | 116 | 584 | 188 | 4 | **1580** |
| Azules | 326 | 38 | 241 | 110 | 3 | **718** |
| Castaños | 343 | 84 | 909 | 412 | 26 | **1774** |
| Oscuros | 98 | 48 | 403 | 681 | 85 | **1315** |
| **Total** | **1455** | **286** | **2137** | **1391** | **118** | **5387** |

## Solución:

Una manera de llegar a una tabla de contingencia I×J es definir **I** variables binarias para las categorías de las filas y **J** para las de las columnas.

**A. Análisis de correspondencias SPSS**

Uno de los objetivos del análisis de correspondencias es describir las relaciones existentes entre dos variables nominales, recogidas en una tabla de correspondencias sobre un espacio de pocas dimensiones, mientras que al mismo tiempo se describen las relaciones entre las categorías de cada variable. Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las categorías, con las categorías similares representadas próximas unas a otras. La proyección de los puntos de una variable sobre el vector, desde el origen hasta un punto de categoría de la otra variable, describe la relación entre ambas variables. El análisis de las tablas de contingencia, a menudo incluye examinar los perfiles de fila y de columna, así como contrastar la independencia a través del estadístico de chi-cuadrado. Sin embargo, el número de perfiles puede ser muy grande y la prueba de chi-cuadrado no revelará la estructura de la dependencia. El procedimiento Tablas de contingencia ofrece varias medidas y pruebas de asociación, pero no puede representar gráficamente alguna relación entre las variables. El **análisis factorial** es una técnica típica para describir las relaciones existentes entre variables en un espacio de pocas dimensiones; sin embargo, el análisis factorial requiere datos de intervalo y el número de observaciones debe ser cinco veces el número de variables. Por su parte, el análisis de correspondencias asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además, el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas.

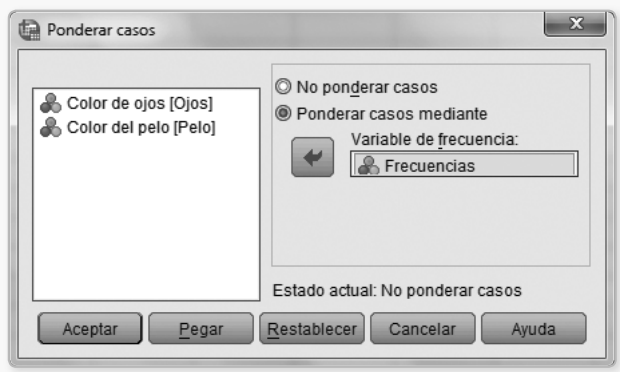
Para resolver el ejemplo anterior, primero debe hallar las frecuencias de las dos variables, tal como se muestra en la tabla siguiente:

| Color de ojos | Color del pelo | Frecuencias |
|---|---|---|
| Claros | Rubio | 688 |
| Azules | Rubio | 326 |
| Castaños | Rubio | 343 |
| Oscuros | Rubio | 98 |
| Claros | Pelirrojo | 116 |
| Azules | Pelirrojo | 38 |
| Castaños | Pelirrojo | 84 |
| Oscuros | Pelirrojo | 48 |
| Claros | Castaño | 584 |
| Azules | Castaño | 241 |
| Castaños | Castaño | 909 |
| Oscuros | Castaño | 403 |
| Claros | Oscuro | 188 |
| Azules | Oscuro | 110 |
| Castaños | Oscuro | 412 |
| Oscuros | Oscuro | 681 |
| Claros | Negro | 4 |
| Azules | Negro | 3 |
| Castaños | Negro | 26 |
| Oscuros | Negro | 85 |

Luego, ingrese los datos al SPSS.

**B. Ejecución del análisis**

Antes de empezar con el análisis de correspondencia, es necesaria la configuración de los datos que requiere que los casos se ponderen mediante la variable Frecuencia. Para ello, elija en los menús Datos > Ponderar casos.



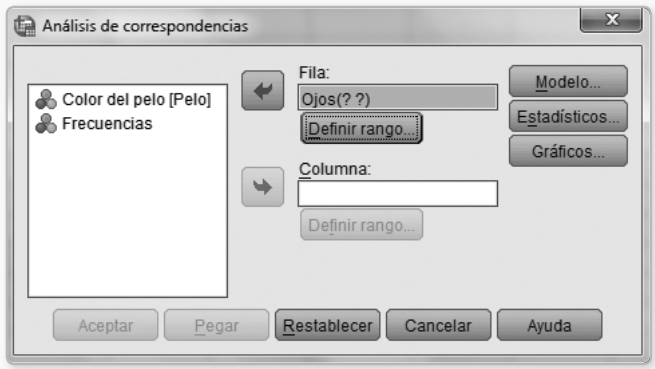

**NOTA**

*Si no pondera los casos, no se podrá realizar análisis de correspondencia.*

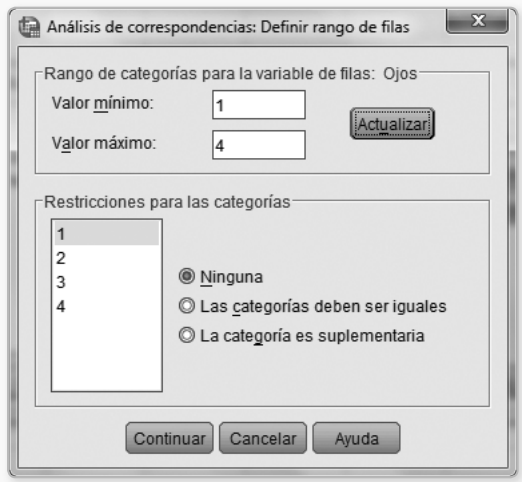
Dentro del menú Analizar se encuentra el procedimiento **Reducción de datos.** El aspecto de dicho menú es el siguiente:

Analizar > Reducción de datos (Data Reduction) > Análisis de correspondencia...

Muestra el cuadro de diálogo Análisis de correspondencias.

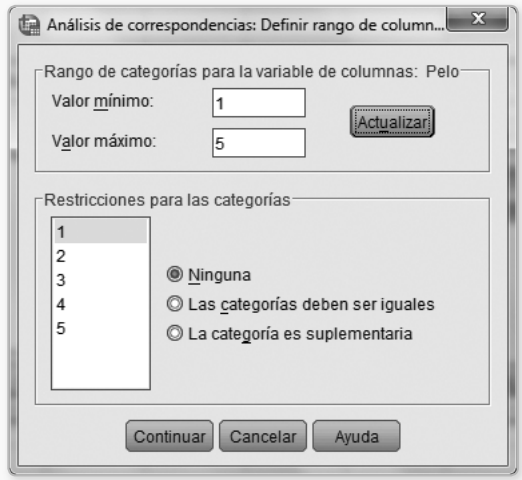



Seleccione *Ojos* como variable de fila y pulse Definir rango: cuadro de diálogo Definir rango de filas.

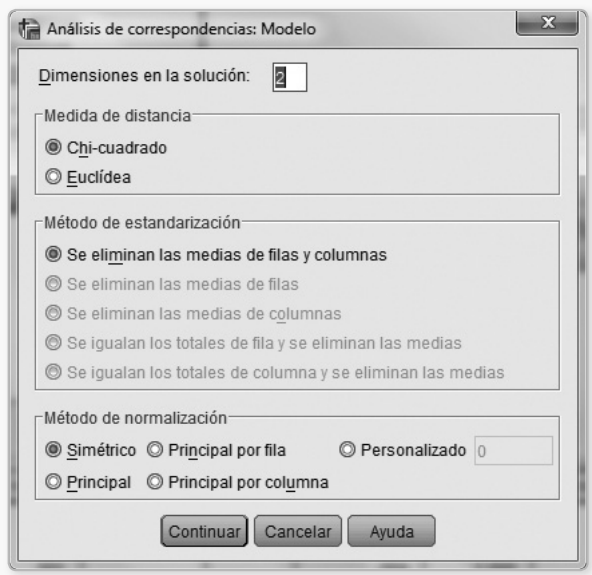



Escriba 1 como valor Mínimo y escriba cuatro como Valor máximo; luego, pulse Actualizar. Finalmente, pulse Continuar.

A continuación, seleccione Pelo como variable Columna, y pulse Definir rango en el cuadro de diálogo Análisis de correspondencias: Definir rango de columnas.




Escriba 1 como valor mínimo, luego escriba cinco como valor máximo y luego pulse Actualizar > Continuar. Pulse Modelo en el cuadro de diálogo Análisis de correspondencias: Modelo.



Seleccione Simétrico como método de normalización y luego pulse en Continuar. A continuación, pulse en Gráficos en el cuadro de diálogo Análisis de correspondencias: Gráficos.




Seleccione Puntos de fila y Puntos de columna en el grupo Diagramas de dispersión; luego pulse en Continuar. Finalmente, pulse en Aceptar en el cuadro de diálogo Análisis de correspondencias, obteniéndose los siguientes resultados:

**Tabla de correspondencias**

| Color de ojos | Color del pelo | | | | | |
|---|---|---|---|---|---|---|
| | Castaño | Negro | Oscuro | Pelirrojo | Rubio | Margen activo |
| Azules | 241 | 3 | 110 | 38 | 326 | 718 |
| Castaños | 909 | 26 | 412 | 84 | 343 | 1774 |
| Claros | 584 | 4 | 188 | 116 | 688 | 1580 |
| Oscuros | 403 | 85 | 681 | 48 | 98 | 1315 |
| Margen activo | 2137 | 118 | 1391 | 286 | 1455 | 5387 |

*Figura 7.4 Tabla de correspondencia bidimensional*

### Dimensionalidad

La inercia por dimensión muestra la descomposición de la inercia total a lo largo de cada dimensión. Una dimension explica el 86.6% de la inercia total. Si se añade una segunda dimensión, solo se añade un 13.1% a la inercia explicada. Por tanto, puede elegir utilizar una representación en dos dimensiones.

**Resumen**

| Dimensión | Valor singular | Inercia | Chi cuadrado | Sig. | Proporción de inercia | | Valor singular de confianza | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Contabilizado para | Acumulado | Desviación estándar | Correlación 2 |
| 1 | ,446 | ,199 | | | ,866 | ,866 | ,012 | ,274 |
| 2 | ,173 | ,030 | | | ,131 | ,996 | ,013 | |
| 3 | ,029 | ,001 | | | ,004 | 1,000 | | |
| Total | | ,230 | 1240,039 | ,000[a] | 1,000 | 1,000 | | |

a. 12 grados de libertad

## Contribuciones

Una visión general de los puntos de fila muestra las contribuciones de estos a la inercia de las dimensiones, y las contribuciones de las dimensiones a la inercia de los puntos de fila. Si todos los puntos contribuyen de igual manera a la inercia, las contribuciones serían 0,25 (1/4): *oscuros y claros* contribuyen en una parte importante a la inercia de la primera dimensión, *castaños* contribuye con mayor cantidad a la inercia de la segunda dimensión y *azules* contribuye muy poco a ambas dimensiones.

**Puntos de fila generales[a]**

| Color de ojos | Masa | Puntuación en dimensión | | Inercia | Contribución | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Del punto en la inercia de dimensión | | De la dimensión en la inercia del punto | | |
| | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| Azules | ,133 | ,599 | -,397 | ,026 | ,107 | ,121 | ,836 | ,143 | ,979 |
| Castaños | ,329 | -,050 | ,588 | ,020 | ,002 | ,657 | ,018 | ,981 | ,999 |
| Claros | ,293 | ,660 | -,212 | ,060 | ,286 | ,076 | ,956 | ,039 | ,995 |
| Oscuros | ,244 | -1,052 | -,322 | ,125 | ,605 | ,145 | ,965 | ,035 | 1,000 |
| Total activo | 1,000 | | | ,230 | 1,000 | 1,000 | | | |

a. Normalización simétrica

La visión general de los puntos de columna muestra las contribuciones que implican a los puntos de columna. El color de pelo oscuro, rubio y negro contribuye sobre todo a la primera dimensión; mientras que castaño explica una gran cantidad de la inercia para la segunda dimensión. Mientras que pelirrojo contribuyen muy poco a ambas dimensiones.

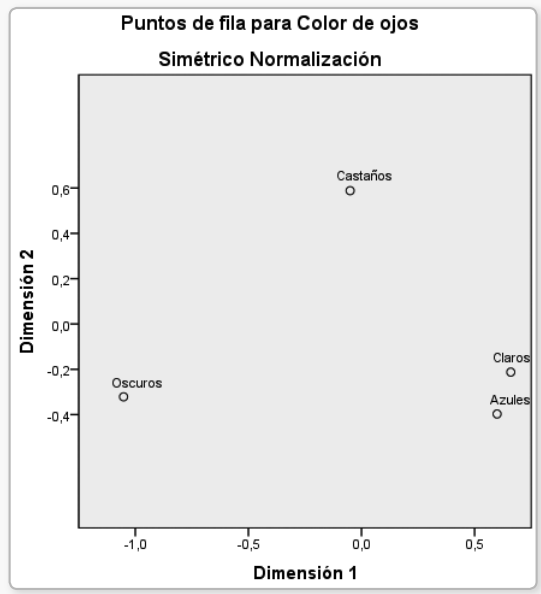
**Puntos de columna generales[a]**

| Color del pelo | Masa | Puntuación en dimensión | | Inercia | Contribución | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Del punto en la inercia de dimensión | | De la dimensión en la inercia del punto | | |
| | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| Castaño | ,397 | ,063 | ,500 | ,018 | ,004 | ,572 | ,039 | ,961 | 1,000 |
| Negro | ,022 | -1,638 | -,688 | ,028 | ,132 | ,060 | ,934 | ,064 | ,998 |
| Oscuro | ,258 | -,881 | -,250 | ,092 | ,449 | ,093 | ,969 | ,030 | 1,000 |
| Pelirrojo | ,053 | ,349 | -,116 | ,004 | ,014 | ,004 | ,770 | ,033 | ,803 |
| Rubio | ,270 | ,814 | -,417 | ,088 | ,401 | ,271 | ,907 | ,093 | 1,000 |
| Total activo | 1,000 | | | ,230 | 1,000 | 1,000 | | | |

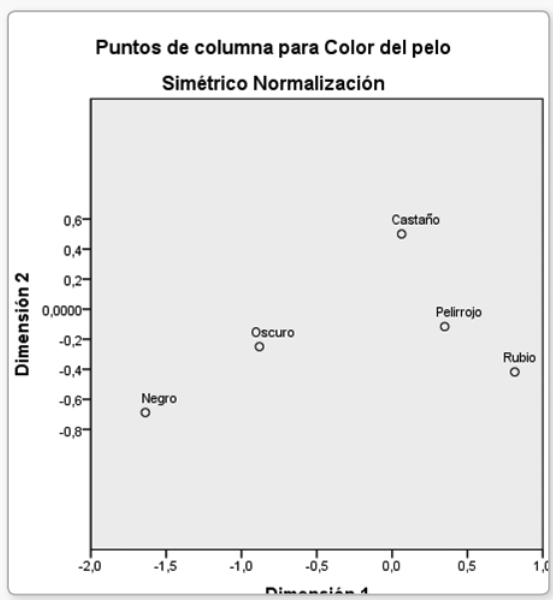
a. Normalización simétrica

**Gráficos**

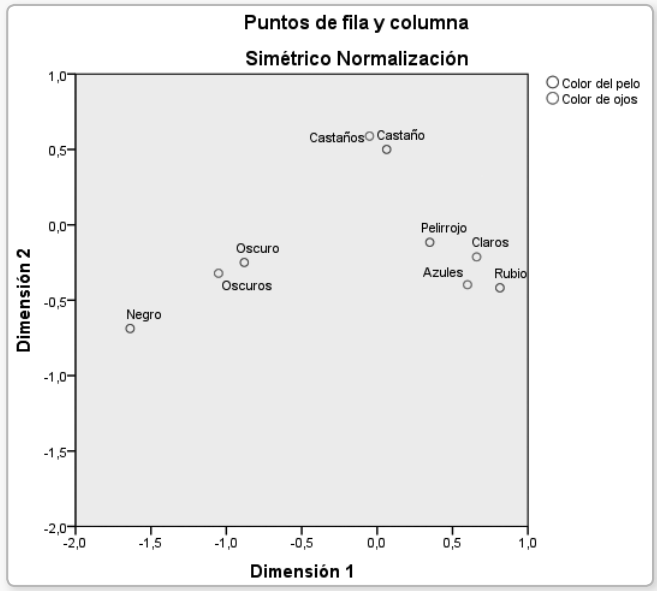
El gráfico de puntos de fila muestra que surgen así tres clasificaciones generales. Situado, en la parte izquierda inferior del gráfico, la primera clasificación representado por *oscuros*; mientras que la segunda clasificación por *claros* y *azules* que son similares entre sí. Finalmente, *castaños*.




Observe, en el gráfico de los puntos de columna, que todas las marcas están lejos entre sí. En consecuencia, los color de cabello, en general, no están relacionas. Sin embargo, pelirrojo con *rubio* tienen una baja correlación.



Finalmente, del gráfico siguiente, se puede concluir que los de color de pelo *pelirrojo* y *rubio* tienen ojos *claros* y *azules*; mientras lo de pelo *castaño* tiene ojos *castaños*; además, color de pelo oscuro y negro tienen color de ojos *oscuros*.



**EJERCICIO PRÁCTICO 2**

El ejercicio permitirá utilizar los principales procedimientos del menú reducción de datos. Los datos usados serán los recopilados en la encuesta del supermercado.

Dentro del menú Analizar se encuentra el procedimiento **Reducción de datos**, el aspecto de dicho menú es el siguiente:

Analizar > Reducción de datos (Data Reduction)



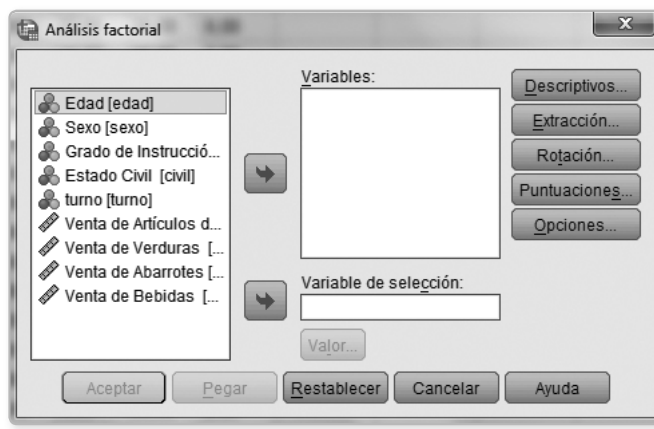

Para realizar el procedimiento estadístico Reducción de datos, se debe seguir los cuadros de diálogo que se presenta. A continuación, se muestra una explicación detallada del cuadro de diálogo correspondiente a los principales procedimientos.

### A. Análisis factorial

El análisis de factorial consta de cuatro fases características:

- El cálculo de una matriz capaz de expresar la variabilidad conjunta de todas las variables.
- La extracción del número óptimo de factores.
- La rotación de la solución para facilitar su interpretación.
- La estimación de las puntuaciones de los sujetos en las nuevas dimensiones.

Para llevar a cabo un análisis de factorial, seleccione la opción **Analizar > Reducción de datos > Análisis factorial**, obteniéndose el cuadro de diálogo siguiente:




El cuadro de la izquierda muestra la lista de todas las variables del archivo (también las que tienen formato cadena; aunque estas solo pueden utilizarse como variables de selección), para llevar a cabo Análisis de factorial con SPSS.

Seleccione el conjunto de variables que se desea analizar y se traslada al cuadro del lado derecho lista de **Variables** (Variables).

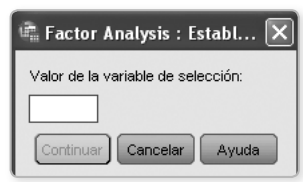



En este ejemplo, las variables Edad, Grado de instrucción, Estado civil, Turno, Venta de verduras, Venta de abarrotes y Venta de bebidas fueron seleccionadas para realizar Análisis factorial.

### B. Variable de selección (Selection Variable).

Este cuadro permite seleccionar unas de las variables del archivo de datos como variable *Filtro* para definir una submuestra de sujetos que cumplen una determinada condición.

Si se traslada la variable al cuadro Variable de selección y pulsa el botón **Valor** (Value...), se obtiene el cuadro de diálogo siguiente:




Introduzca en el cuadro el valor de la variable de selección que identifica a los casos que se desea incluir en el análisis. En el ejemplo, luego de seleccionar las variables mencionadas, acepte (OK) el Visor de resultados muestra los siguientes cuadros.

El cuadro siguiente contiene las comunalidades (Communalities) asignadas inicialmente a las variables (initial) y las comunalididades reproducidas por la solución factorial (Extraction). La **comunalidad de una variable** es la proporción de su varianza que puede ser explicada por el modelo factorial obtenido.

**Communalities**

| | Initial | Extraction |
|---|---|---|
| Edad | 1.000 | .911 |
| Grado de Instrucción | 1.000 | .802 |
| Estado Civil | 1.000 | .702 |
| turno | 1.000 | .903 |
| Venta de Verduras | 1.000 | .452 |
| Venta de Abarrotes | 1.000 | .739 |
| Venta de Bebidas | 1.000 | .722 |

Extraction Method: Principal Component Analysis.

En el ejemplo, la variable Venta de verduras es la peor explicada: el modelo solo es capaz de reproducir el 45.2% de su variabilidad original. Para la solución factorial, se ha utilizado el método de extracción de componentes principales. También, se puede plantear en este momento si dando por bueno el número de factores extraídos, alguna de las variables incluidas podría quedar fuera del análisis.

El cuadro siguiente muestra la tabla *Porcentajes de varianza explicada* se ofrece un listado de los autovalores (Eigenvalues) de la matriz varianza-covarianza y del porcentaje de varianza que representa cada uno de ellos. Los autovalores expresan la cantidad de la varianza total que está explicada por cada factor. Por defecto, se extraen tantos factores como autovalores mayores que uno tiene la matriz analizada.

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2.513 | 35.894 | 35.894 | 2.513 | 35.894 | 35.894 |
| 2 | 1.716 | 24.519 | 60.413 | 1.716 | 24.519 | 60.413 |
| 3 | 1.001 | 14.297 | 74.710 | 1.001 | 14.297 | 74.710 |
| 4 | .744 | 10.622 | 85.332 | | | |
| 5 | .584 | 8.346 | 93.678 | | | |
| 6 | .414 | 5.912 | 99.591 | | | |
| 7 | 2.864E-02 | .409 | 100.000 | | | |

Extraction Method: Principal Component Analysis.

En el ejemplo, hay tres autovalores mayores que uno; por lo que el procedimiento extrae tres factores que consiguen explicar el 74.71% de la varianza de los datos originales. La tabla también muestra para cada factor con autovalor mayor que uno, la suma de las saturaciones al cuadrado.
La información de esta tabla puede utilizarse para tomar una decisión sobre el número idóneo de factores que deben extraerse. Por ejemplo, un mínimo del 85% de la variabilidad contenida en los datos sería necesario obtener cuatro factores.

El cuadro siguiente muestra la tabla donde se encuentra la solución factorial propiamente dicha. Contiene las correlaciones entre las variables originales (o saturaciones) y cada uno de los factores. Esta matriz se llama «matriz de componentes» porque el método usa componentes principales.

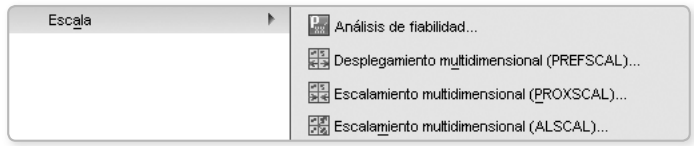
**Component Matrix[a]**

| | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Edad | .938 | .175 | .031 |
| Grado de Instrucción | .010 | .713 | .541 |
| Estado Civil | .325 | .585 | -.504 |
| turno | .935 | .166 | .032 |
| Venta de Verduras | -.650 | .140 | -.098 |
| Venta de Abarrotes | .297 | -.576 | .565 |
| Venta de Bebidas | -.377 | .675 | .351 |

Extraction Method: Principal Component Analysis.
a. 3 components extracted.

En el ejemplo, comparando las saturaciones relativas de cada variable en cada uno de los tres factores, se observa que el primer factor está constituido por las variables Edad (0.938) y Turno (0.935); el segundo factor está constituido por las variables Grado de instrucción, Estado civil, Venta de verduras y Venta de bebidas y tercer factor está constituido por la variable Venta de abarrotes.

## 7.3. ESCALAS




### 7.3.1. ANÁLISIS DE FIABILIDAD

Permite estudiar las propiedades de las escalas de medición y de los elementos que las constituyen. Calcula un número de medidas de fiabilidad de escala que se usan normalmente y también proporciona información sobre las relaciones entre elementos individuales de la escala.

Se pueden emplear los coeficientes de correlación intraclase para calcular estimaciones de la fiabilidad interevaluadores, modelos de fiabilidad disponibles:

- **Alfa (Cronbach).** Modelo de consistencia interna basado en la correlación interelementos promedio.
- **Dos mitades.** Divide la escala en dos partes y examina la correlación entre dichas partes.
- **Guttman.** Calcula los límites inferiores de Guttman para la fiabilidad verdadera.
- **Paralelo.** Asume que todos los elementos tienen varianzas iguales y varianzas error iguales a través de las réplicas.
- **Paralelo estricto.** Asume los supuestos del modelo paralelo y también asume que las medias son iguales a través de los elementos.

### 7.3.2. ESCALAMIENTO MULTIDIMENSIONAL

Trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos; ello se logra asignando las observaciones a posiciones específicas en un espacio conceptual (normalmente de dos o tres dimensiones); de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades dadas.

En muchos casos, las dimensiones de este espacio conceptual son interpretables y se pueden utilizar para comprender mejor los datos. Si las variables se han medido objetivamente, se puede emplear como técnica de reducción de datos (permitirá calcular las distancias, a partir de los datos multivariados si es necesario). El escalamiento multidimensional puede también aplicarse a valoraciones subjetivas de disimilaridad entre objetos o conceptos. Además, puede tratar datos de disimilaridad procedentes de múltiples fuentes, como podrían ser múltiples evaluadores o múltiples sujetos evaluados por un cuestionario.

### 7.3.3. ESCALAMIENTO MULTIDIMENSIONAL (PROXSCAL)

Trata de encontrar la estructura existente en un conjunto de medidas de proximidades entre objetos. Esto se logra asignando las observaciones a posiciones específicas, en un espacio conceptual de pocas dimensiones; de modo que las distancias entre los puntos en el espacio concuerden al máximo con las similaridades (o disimilaridades) dadas. El resultado es una representación de mínimos cuadrados de los objetos en dicho espacio de pocas dimensiones que, en muchos casos, ayuda a entender mejor los datos.

CAPÍTULO

8

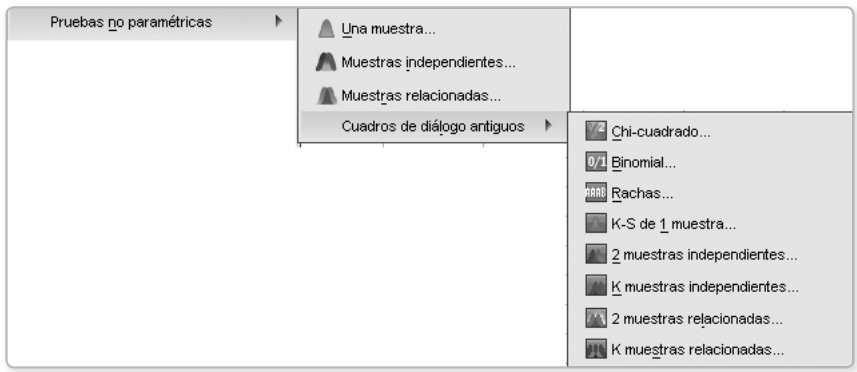
# PRUEBAS NO PARAMÉTRICAS

Este es un capítulo práctico, donde se presentan numerosas técnicas estadísticas no paramétricas básicas, se discuten situaciones en las que son aplicables cada una de ellas y las hipótesis previas necesarias para realizar e interpretar los resultados.

Los contrastes no paramétricos que se presentan en este capítulo son: pruebas de los signos (binomial), dos muestras independientes (prueba de Mann-Whitney), dos muestras relacionadas (prueba de Wilcoxon), más de dos muestras independientes (prueba de Kruskal Wallis), más de dos muestras relacionadas (prueba de Friedman), k muestras relacionadas,contraste sobre una proporción, dos proporciones relacionadas, más de dos proporciones relacionadas, dos proporciones independientes, prueba $X^2$ de Pearson, $X_2$ sobre bondad de ajuste y tablas de contingencia.

## 8.1. PRUEBAS NO PARAMÉTRICAS

Este tema ofrece una descripción de las técnicas de análisis que el SPSS clasifica como pruebas no paramétricas. Todas ellas pueden considerarse no paramétricas, ya que no plantean hipótesis sobre parámetros. Todas estas pruebas se encuentran en la opción Pruebas no paramétricas, del menú Analizar, y son las siguientes:




- **Chi-cuadrado.** Tabula una variable en categorías y calcula un estadístico de chi-cuadrado. Esta prueba de bondad de ajuste compara las frecuencias observadas y esperadas en cada categoría, para contrastar si todas contienen la misma proporción de valores, o si cada categoría contiene una proporción de valores especificada por el usuario.
- **Binomial.** Compara las frecuencias observadas de las dos categorías de una variable dicotómica, con las frecuencias esperadas en una distribución binomial con un parámetro de probabilidad especificado. Por defecto, el parámetro de probabilidad para ambos grupos es 0,5. Para cambiar las probabilidades, puede introducirse una proporción de prueba para el primer grupo. La probabilidad del segundo grupo será 1, menos la probabilidad especificada para el primer grupo.
- **Rachas.** Contrasta si el orden de aparición de dos valores de una variable es aleatorio. Una **racha** es una secuencia de observaciones similares. Una muestra con un número excesivamente grande o pequeño de rachas sugiere que la muestra no es aleatoria.
- **K-S de 1 muestra.** La prueba de Kolmogorov-Smirnov para una muestra, compara la función de distribución acumulada observada de una variable, con una distribución teórica determinada que puede ser la normal, la uniforme, la de Poisson o la exponencial. La Z de Kolmogorov-Smirnov se calcula a partir de la diferencia mayor (en valor absoluto), entre las funciones de distribución acumuladas teórica y observada. Esta prueba de bondad de ajuste contrasta si las observaciones pueden proceder de la distribución especificada de manera razonable.
- **Dos muestras independientes:** Compara dos grupos de casos existentes en una variable.
- **K muestras independientes:** Compara dos o más grupos de casos, respecto a una variable.
- **Dos muestras relacionadas:** Compara las distribuciones de dos variables.
- **K muestras relacionadas:** Compara las distribuciones de dos o más variables.

## 8.2. EJERCICIO PRÁCTICO 1

Repase los procedimientos básicos de las pruebas no paramétricas. Se trabajará con la información presentada a continuación. Los datos aparecen recopilados en las variables *Día1*, *Día2*, *Día3* que indican el número de horas trabajadas en los tres días; además, se han recopilado la variable *Sexo* que toma valores 0 para mujeres y 1 para hombres, y *Edad*, 1 para adolescentes, 2 para jóvenes y 3 para adultos.

Finalmente, se incluyen tres datos relacionados con las terapias que ha recibido el sujeto. Todas ellas toman 1, si el sujeto ha recibido la terapia, y 0 en caso contrario. Las variables son *Estrés*, *Ansiedad* y *Fobia*.

Los datos representan 10 sujetos que participaron en el estudio, cuyo resultado se muestra en la tabla siguiente:

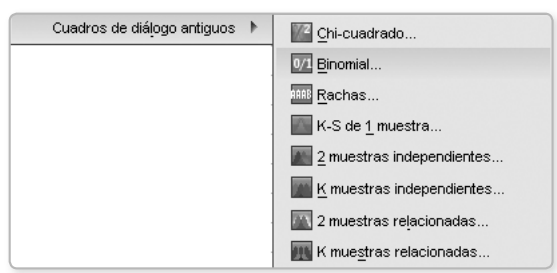
| Sujeto | Día1 | Día2 | Día3 | Sexo | Edad | Estrés | Ansiedad | Fobia |
|---|---|---|---|---|---|---|---|---|
| 1 | 5.7 | 6.3 | 6.8 | 0 | 1 | 0 | 1 | 0 |
| 2 | 4.8 | 5.4 | 4.3 | 1 | 3 | 0 | 0 | 0 |
| 3 | 7.6 | 7.9 | 8.6 | 0 | 2 | 0 | 0 | 0 |
| 4 | 5.7 | 6.0 | 6.7 | 1 | 1 | 0 | 0 | 1 |
| 5 | 3.8 | 4.6 | 4.9 | 1 | 3 | 0 | 0 | 0 |
| 6 | 7.5 | 8.0 | 8.5 | 1 | 2 | 0 | 1 | 0 |
| 7 | 6.4 | 7.0 | 7.3 | 1 | 1 | 1 | 1 | 0 |
| 8 | 7.7 | 8.1 | 8.5 | 0 | 2 | 0 | 0 | 0 |
| 9 | 4.0 | 4.5 | 5.0 | 0 | 3 | 1 | 0 | 0 |
| 10 | 5.7 | 6.1 | 6.8 | 1 | 2 | 1 | 0 | 1 |

## 8.3. CONTRASTES NO PARAMÉTRICOS

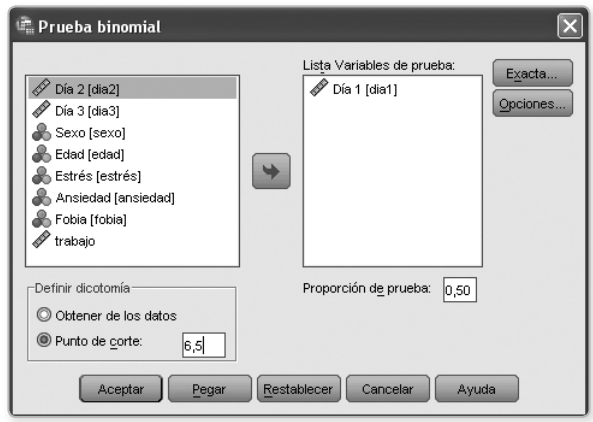
Los contrastes no paramétricos se ejecutan mediante la opción del menú Analizar > Pruebas no paramétricas.

### 8.3.1. PRUEBAS DE LOS SIGNOS (BINOMIAL)

En SPSS, se ejecuta mediante la opción **Prueba binomial**.



Por ejemplo, suponga que se desea contrastar la hipótesis acerca de que la mediana de la variable *Día1* es 6,5. Esto se indica del siguiente modo en el cuadro de diálogo Prueba binomial.



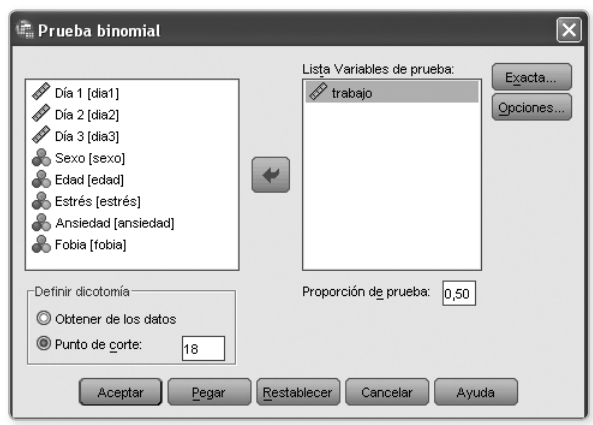

En la casilla denominada Punto de corte, se ha introducido el valor para la mediana en la hipótesis nula; en Contrastar proporción, el valor 0,5 que es la proporción de sujetos con puntuaciones menores o iguales a 6,5, según la hipótesis nula. La salida de resultados se basa en el estadístico de contraste:

**Prueba binomial**

| | | Categoría | N | Prop. observada | Prop. de prueba | Significación exacta (bilateral) |
|---|---|---|---|---|---|---|
| Día 1 | Grupo 1 | <= 6,5 | 7 | ,70 | ,50 | ,344 |
| | Grupo 2 | > 6,5 | 3 | ,30 | | |
| | Total | | 10 | 1,00 | | |

En el ejemplo, la tabla indica que 7 sujetos obtuvieron puntuaciones iguales o menores que 6,5. Esto se corresponde con el 70% de los sujetos de la muestra, mientras que el porcentaje esperado bajo $H_0$ es 50%. El valor del nivel crítico bilateral es 0,344, y se utiliza para tomar una decisión sobre $H_0$.

La prueba binomial también permite realizar contrastes sobre cualquier otro centil, solo se necesita cambiar la proporción introducida en la casilla Contrastar proporción. Por ejemplo, para contrastar la hipótesis de que el centil 90 de la variable Trabajo (Día 1+Día 2+Día 3) es el valor 18, y se indica de la siguiente manera:



La salida de resultados se basa en el estadístico de contraste.

**Prueba binomial**

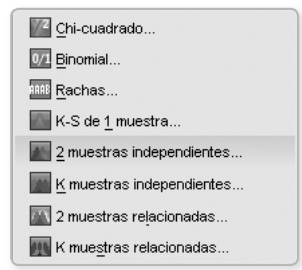
| | | Categoría | N | Prop. observada | Prop. de prueba | Significación exacta (bilateral) |
|---|---|---|---|---|---|---|
| trabajo | Grupo 1 | <= 18 | 3 | ,30 | ,50 | ,344 |
| | Grupo 2 | > 18 | 7 | ,70 | | |
| | Total | | 10 | 1,00 | | |

En el ejemplo, la tabla indica que tres sujetos obtuvieron puntuaciones iguales o menores que 18. Esto se corresponde con el 30% de los sujetos de la muestra; mientras que el porcentaje esperado bajo $H_0$ es el 90%. El valor del nivel crítico bilateral es 0,00 y se utiliza para tomar una decisión sobre $H_0$.

La opción Definir la dicotomía especifica que se va a comprobar si el 90% de los sujetos tienen valores iguales o menores que 18.

## 8.3.2. CONTRASTES PARA DOS MUESTRAS

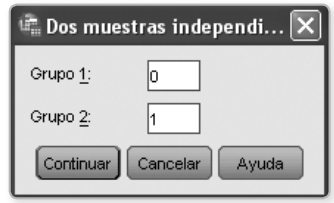
### A. Dos muestras independientes: prueba de Mann-Whitney




Seleccionando los contrastes **2 Muestras independientes**, se muestra un cuadro de diálogo en el que se especifican las variables y el tipo de contraste. Para contrastar la hipótesis de que el valor esperado de *Día2* es igual en los dos grupos de la variable *Estrés,* se indica lo siguiente:



Mediante el botón Definir grupos, es necesario indicar que los dos grupos de la variable **Estrés** son los correspondientes a los valores 0 y 1.




La salida de resultados muestra distintas tablas. Una de ellas contiene la suma de los rangos asignados a las puntuaciones en *Día2* en los dos grupos de *Estrés*.

**Ranks**

| | Estrés | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| Día 2 | No | 7 | 6.00 | 42.00 |
| | Si | 3 | 4.33 | 13.00 |
| | Total | 10 | | |

En la tabla separada, aparecen el estadístico de contraste U de Mann – Whitney y el nivel crítico.

**Test Statistics[b]**

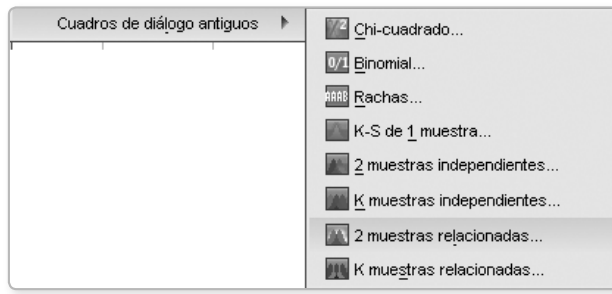
| | Día 2 |
|---|---|
| Mann-Whitney U | 7.000 |
| Wilcoxon W | 13.000 |
| Z | -.798 |
| Asymp. Sig. (2-tailed) | .425 |
| Exact Sig. [2*(1-tailed Sig.)] | .517[a] |

a. Not corrected for ties.

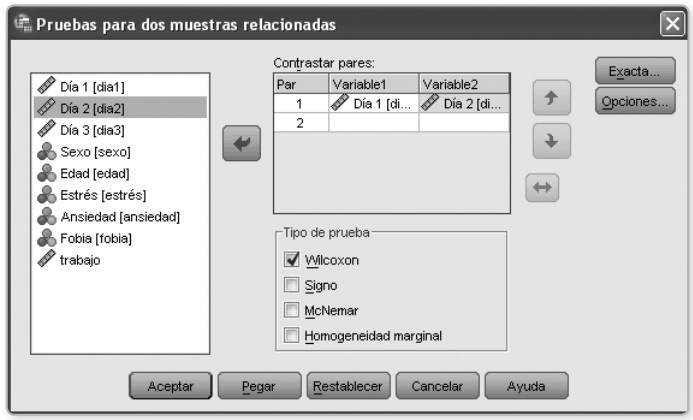
b. Grouping Variable: Estrés

### B. Dos muestras relacionadas: Prueba de Wilcoxon

Se encuentran en la opción **2 Muestras relacionadas**.



Por ejemplo, los sujetos del grupo *Día1* son los mismos que los de *Día 2*, por lo que ambas son muestras relacionadas. Para comprobar que sus medianas son iguales, se indica lo siguiente:




En las tablas siguientes se muestra el estadístico de contraste.

**Ranks**

| | | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| Día 2 - Día 1 | Negative Ranks | 0[a] | .00 | .00 |
| | Positive Ranks | 10[b] | 5.50 | 55.00 |
| | Ties | 0[c] | | |
| | Total | 10 | | |

a. Día 2 < Día 1
b. Día 2 > Día 1
c. Día 1 = Día 2

**Test Statistics[b]**

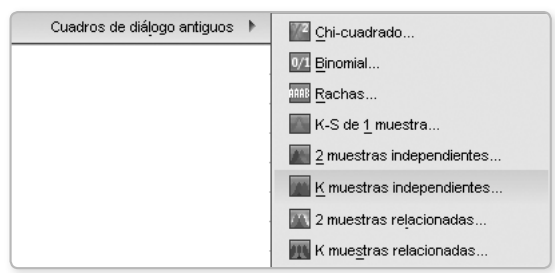
| | Día 2 - Día 1 |
|---|---|
| Z | -2.816[a] |
| Asymp. Sig. (2-tailed) | .005 |

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

### 8.3.3. CONTRASTES PARA MÁS DE DOS MUESTRAS

#### A. Más de dos muestras independientes: Prueba de Kruskal Wallis

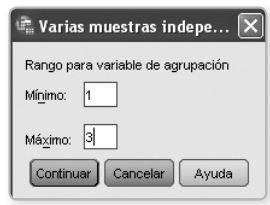
Se especifica mediante la opción K muestras independientes...



A continuación, puede verse cómo contrastar la hipótesis de que la mediana de *Día 3* es la misma en los tres grupos de edad.




Mediante Definir rango, se ha indicado que los grupos de *Edad* vienen definidos por los valores 1 a 3.




La salida de resultados muestra la suma de rangos en cada grupo de *Edad* y el valor del estadístico $X^2$.

**Ranks**

| | Edad | N | Mean Rank |
|---|---|---|---|
| Día 3 | Adolescente | 3 | 5.50 |
| | Jóven | 4 | 8.13 |
| | Adulto | 3 | 2.00 |
| | Total | 10 | |

**Test Statistics[a,b]**

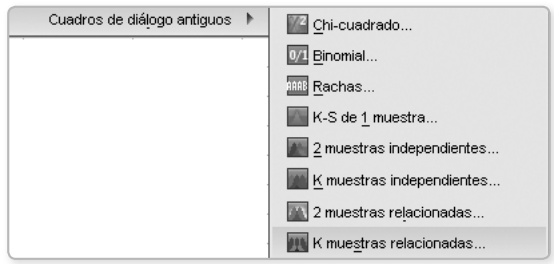
| | Día 3 |
|---|---|
| Chi-Square | 7.102 |
| df | 2 |
| Asymp. Sig. | .029 |

a. Kruskal Wallis Test

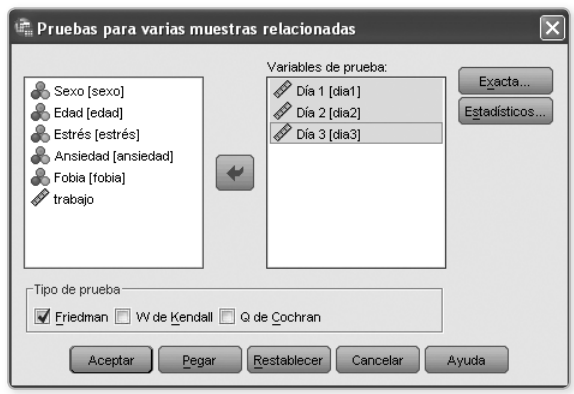
b. Grouping Variable: Edad

### B. Más de dos muestras relacionadas: Prueba de Friedman

Se especifica mediante la opción **K muestras relacionadas**.




En el ejemplo, las variables de los tres *Días* están formadas por los mismos sujetos, por lo que son relacionadas. Del siguiente modo, se comprueba si sus medianas son iguales:




La salida de resultados muestra el número de casos estadístico de contraste X, los grados de libertad y el nivel crítico.

**Ranks**

| | Mean Rank |
|---|---|
| Día 1 | 1.10 |
| Día 2 | 2.10 |
| Día 3 | 2.80 |

**Test Statistics[a]**

| | |
|---|---|
| N | 10 |
| Chi-Square | 14.600 |
| df | 2 |
| Asymp. Sig. | .001 |

a. Friedman Test

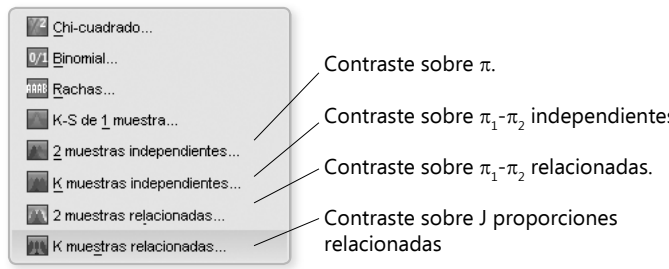
El nivel crítico asociado al estadístico $X^2$ permite rechazar la hipótesis nula de igualdad de medianas en las tres variables.

## 8.4. EJERCICIO PRÁCTICO 2

Se repasarán los procedimientos básicos de las pruebas no paramétricas sobre proporciones. Se trabajará con la información del *Ejercicio práctico 1*.

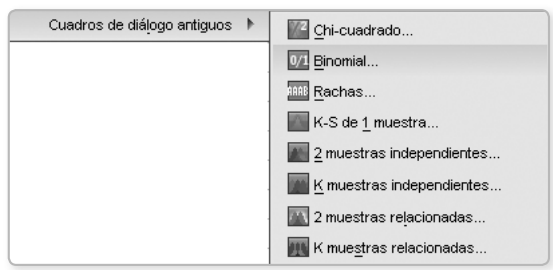
### A. Contrastes sobre proporciones

Los procedimientos para realizar contrastes sobre proporciones se encuentran dentro del menú **Analizar** > **Pruebas no paramétricas**. El procedimiento particular varía dependiendo del número de grupos, y si estos se encuentran relacionados.

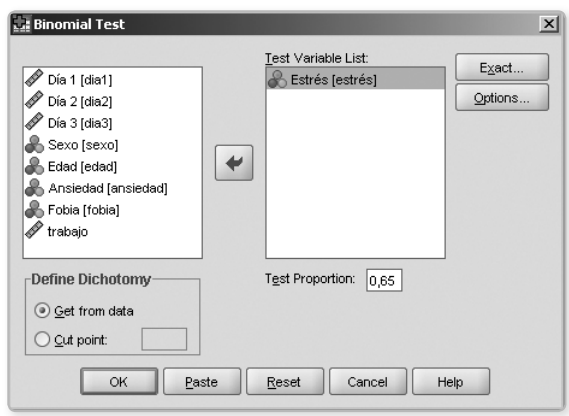



### B. Contraste sobre una proporción

Este contraste es equivalente a la prueba no paramétrica binomial, vista en la práctica anterior.




Por ejemplo, para contrastar la hipótesis de que la proporción de sujetos que reciben la terapia de estrés es 0,65, se indica que:



Luego, se pulsa Aceptar (OK) y se obtiene el resultado de la tabla.

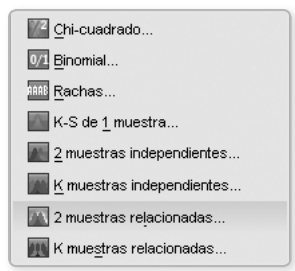
**Binomial Test**

| | | Category | N | Observed Prop. | Test Prop. | Exact Sig. (1-tailed) |
|---|---|---|---|---|---|---|
| Estrés | Group 1 | No | 7 | .70 | .65 | .514 |
| | Group 2 | Si | 3 | .30 | | |
| | Total | | 10 | 1.00 | | |

El nivel crítico unilateral se muestra en la última columna de la tabla de resultados. Usando $\alpha$ = 0,05 no puede rechazarse la hipótesis de que la verdadera proporción de sujetos que recibe la terapia de estrés es 0,65.

### C. Dos proporciones relacionadas

La prueba de McNemar se encuentra disponible en el procedimiento **Analizar > Pruebas no paramétricas > 2 muestras relacionadas**.



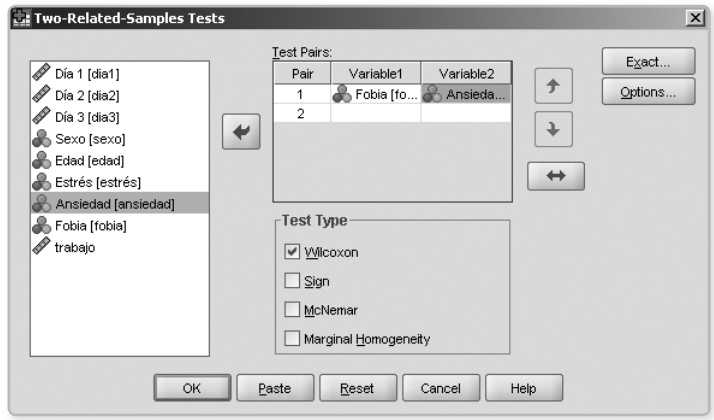

Por ejemplo, se sabe que los sujetos de esta muestra pueden haber recibido terapia contra el estrés, así como otras dos terapias contra situaciones de ansiedad y fobias. Se desea comprobar si los sujetos que reciben la segunda terapia es más probable que reciban también la tercera. Esto indicaría que los individuos que presentan un trastorno tienen mayor probabilidad de presentar el otro.

El contraste se realiza mediante el procedimiento **Analizar > Pruebas no paramétricas > 2 muestras relacionadas**. El cuadro de diálogo de este procedimiento permite escoger distintas pruebas; en concreto, para realizar la prueba de McNemar sobre las variables *Ansiedad* y *Fobia* indicadas, si selecciona las dos variables, *Ansiedad* y *Fobia*. Se desea estudiar si la proporción de sujetos en ambas es igual o no; entonces, la prueba de hipótesis estaría dada por:

$$H_0 : \pi_1 - \pi_2 = 0$$

$$H_1 : \pi_1 - \pi_2 \pi 0$$

Señale la prueba de Mc Nemar. Esta prueba es la que contrasta hipótesis sobre dos proporciones relacionadas.




El Visor de resultados muestra la tabla de contingencia de ambas variables y los estadísticos de contraste.

**Ansiedad & Fobia**

| Ansiedad | Fobia 0 | Fobia 1 |
|---|---|---|
| 0 | 5 | 2 |
| 1 | 3 | 0 |

**Test Statistics[b]**

| | Ansiedad & Fobia |
|---|---|
| N | 10 |
| Exact Sig. (2-tailed) | 1.000[a] |

a. Binomial distribution used.

b. McNemar Test

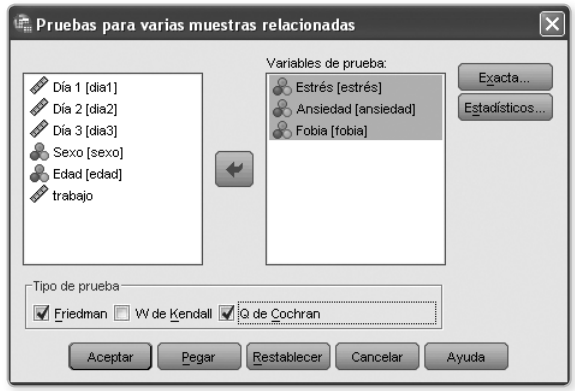
El resultado del test de la hipótesis. Como $\rho = 1.000^a > \alpha$, se mantiene $H_0$; es decir, no existe diferencia significativa entre la proporción de sujetos que reciben las terapias de *Ansiedad* y *Fobia.*

### D. Más de dos proporciones relacionadas

Para realizar la prueba de Cochran con el SPSS, se escoge la opción **Analizar > Pruebas no paramétricas > K muestras relacionadas**.



Por ejemplo, si quisiese contrastarse la hipótesis de que la proporción de sujetos que reciben alguna de las tres terapias es la misma, entonces:




La salida de resultados muestra la distribución de frecuencias de cada variable y el estadístico de contraste *Q*:

**Frequencies**

| | Value | |
|---|---|---|
| | 0 | 1 |
| Estrés | 7 | 3 |
| Ansiedad | 7 | 3 |
| Fobia | 8 | 2 |

**Test Statistics**

| | |
|---|---|
| N | 10 |
| Cochran's Q | .333[a] |
| df | 2 |
| Asymp. Sig. | .846 |

a. 0 is treated as a success.

Resultado del test de la hipótesis: Como $\rho = 0.846 > \alpha$, se mantiene $H_0$; es decir, no existe diferencia significativa entre la proporción de sujetos que reciben las terapias de *Estrés, Ansiedad* y *Fobia.*

### E. Dos proporciones independientes

Este tipo de contraste se realiza mediante la prueba $X^2$ sobre igualdad de proporciones, que se comenta en la siguiente práctica.

## 8.5. PRUEBA $X^2$ DE PEARSON

### 8.5.1. $X^2$ SOBRE BONDAD DE AJUSTE

Para comprobar el ajuste de la distribución de frecuencias de una variable discreta a una distribución teórica, se utiliza la prueba X, que se encuentra disponible en **Analizar** > **Pruebas no paramétricas** > **Prueba chi-cuadrado**.




A modo de ejemplo, se va a comprobar el ajuste de la variable *Edad* a una distribución uniforme. Utilizando el procedimiento **Frecuencias**, puede obtenerse la distribución de frecuencias de esta variable ($H_0$: $\pi_1 = \pi_2 = \pi_3$).

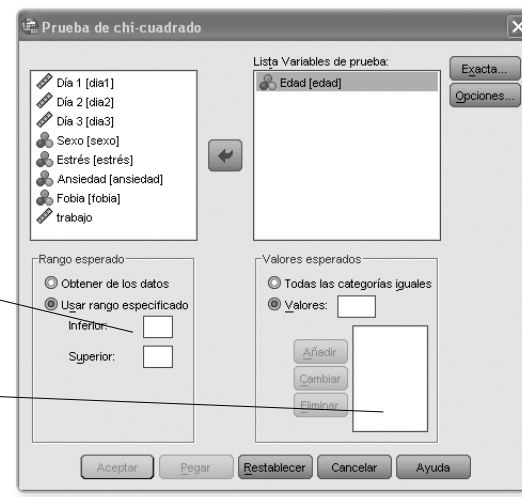
**Edad**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Adolescente | 3 | 30.0 | 30.0 | 30.0 |
| | Jóven | 4 | 40.0 | 40.0 | 70.0 |
| | Adulto | 3 | 30.0 | 30.0 | 100.0 |
| | Total | 10 | 100.0 | 100.0 | |

Se observa que hay tres sujetos adolescentes y adultos, y cuatro sujetos jóvenes de la variable *Edad*. La distribución uniforme asume que la frecuencia de sujetos es la misma en todos los niveles de edad. Para comprobar si la desviación respecto a este supuesto es estadísticamente significativa, se emplea el procedimiento de chi-cuadrado. Al escoger el procedimiento **Analizar** > **Pruebas no paramétricas** > **Prueba Chi-cuadrado**, se muestra el cuadro de diálogo de la prueba $X^2$. En el ejemplo, se usa la variable *Edad* luego:

Aquí se indica la categoría de la variable. Por ejemplo: 0 y1.

Aquí se indica la frecuencia: mi - nπi para cada categoría. Por ejemplo: adolescentes = 10*0.3=3; adultos = 10*0.2 = 2 y jóvenes =10*0.5 = 5.
En este ejemplo, el modelo a contrastar es el uniforme; entonces, se seleccionan todas las categorías iguales.



En el ejemplo, se selecciona la variable *Edad* y la opción Todas las categorías iguales, que indican que la frecuencia esperada de todos los valores de *Edad* es la misma y, por tanto, que se está comprobando el ajuste a una distribución uniforme.

$$H_0\ f(x) = M\ (n = 10\ y\ \pi = 1/3)$$

La salida de resultados muestra una tabla con las frecuencias esperadas, observadas y las residuales (la diferencia entre ambas), y otra con el valor del estadístico de contraste $X^2$, los grados de libertad y el nivel crítico.

**Edad**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| Adolescente | 3 | 3.3 | -.3 |
| Jóven | 4 | 3.3 | .7 |
| Adulto | 3 | 3.3 | -.3 |
| Total | 10 | | |

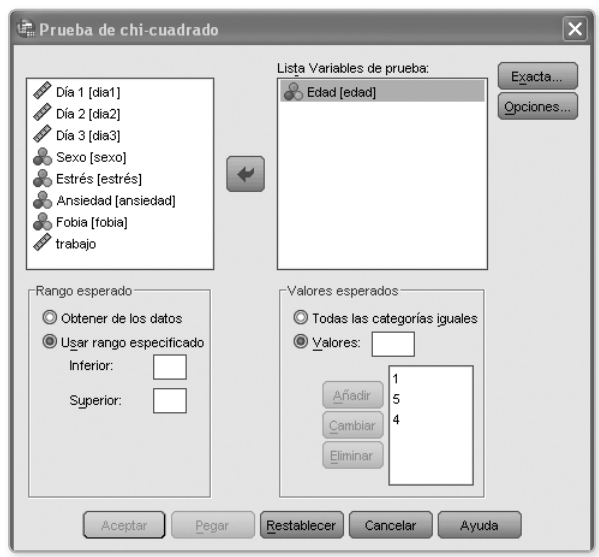
**Test Statistics**

| | Edad |
|---|---|
| Chi-Square[a] | .200 |
| df | 2 |
| Asymp. Sig. | .905 |

a. 3 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.3.

Este procedimiento también permite comprobar el ajuste a cualquier otra distribución discreta. Para ello, se debe indicar directamente la frecuencia esperada para cada valor. Por ejemplo, si se quiere comprobar la frecuencia esperada para cada valor: si se quiere comprobar que la frecuencia esperada de los niveles 1, 2 y 3 de *Edad* es 4, 5 y 1, respectivamente; es decir:

$$H_0\ f(x) = M(n=10,\ \pi_1 = 0.40,\ \pi_2 = 0.50,\ \pi_3 = 0.10)$$



Se obtienen los siguientes resultados:

**Edad**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| Adolescente | 3 | 4.0 | -1.0 |
| Jóven | 4 | 5.0 | -1.0 |
| Adulto | 3 | 1.0 | 2.0 |
| Total | 10 | | |

**Test Statistics**

| | Edad |
|---|---|
| Chi-Square[a] | 4.450 |
| df | 2 |
| Asymp. Sig. | .108 |

a. 2 cells (66.7%) have expected frequencies less than 5. The minimum expected cell frequency is 1.0.

## 8.6. TABLAS DE CONTINGENCIA

Este tipo de contraste también se puede realizar mediante la prueba $X^2$ de las tablas de contingencia, las cuales contienen la distribución conjunta de dos o más variables, y se han estudiado en relación con los contrastes sobre independencia e igualdad de proporciones. En SPSS, se pueden llevar a cabo por medio del procedimiento **Analizar > Estadísticos descriptivos > Tablas de contingencia**. Se va a obtener la tabla de contingencia de las variables *Estrés* y *Edad*. Esto se indica de la siguiente manera:




La salida de resultados de SPSS muestra la siguiente tabla:

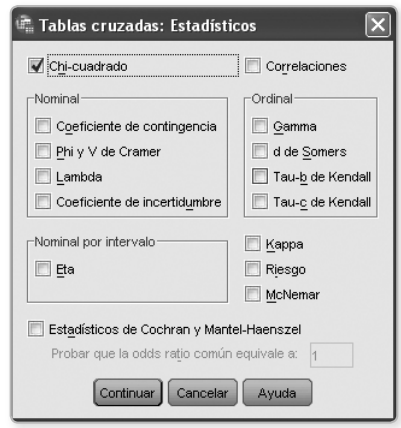
**Estrés * Edad Crosstabulation**

Count

| | | Edad | | | Total |
|---|---|---|---|---|---|
| | | Adolescente | Jóven | Adulto | |
| Estrés | No | 2 | 3 | 2 | 7 |
| | Si | 1 | 1 | 1 | 3 |
| Total | | 3 | 4 | 3 | 10 |

Cada casilla contiene el número observado de sujetos que obtienen un determinado par de valores en *Estrés* y *Edad*. Por ejemplo, hay dos sujetos, el grupo adolescentes que reciben la terapia de estrés, el número total de sujetos en la categoría Adulto es tres, etc.

Para calcular el valor de $X^2$ (y contrastar la hipótesis $H_0$: X y Y son independientes), se pulsa el botón Estadísticos del cuadro de diálogo Procedimiento tablas de contingencia, cuyo aspecto es el siguiente:




Este ejemplo se puede interpretar también como un contraste de igualdad de proporciones, en el que se compara la distribución de la variable *Estrés* (variable dicotómica) en cada grupo de *Edad* (J = 3). El valor del estadístico aparece del siguiente modo en los resultados:

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | .079[a] | 2 | .961 |
| Likelihood Ratio | .080 | 2 | .961 |
| Linear-by-Linear Association | .000 | 1 | 1.000 |
| N of Valid Cases | 10 | | |

a. 6 cells (100.0%) have expected count less than 5. The minimum expected count is .90.

El valor de X está en la primera línea del cuadro, junto con los grados de libertad y el nivel crítico [P(X 0,833) =0,361]. Los índices de asociación basados en chi-cuadrado se muestran en otra tabla diferente. En este caso, puesto que las variables son independientes, estos índices adoptan valores muy pequeños.

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | .089 | .961 |
| | Cramer's V | .089 | .961 |
| | Contingency Coefficient | .089 | .961 |
| N of Valid Cases | | 10 | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Es posible obtener tablas de contingencia con más de dos variables, utilizando la casilla inferior del cuadro de diálogo. Por ejemplo, del siguiente modo se especifica una tabla de contingencia con las variables *Edad, Estrés y Sexo.*




La salida de resultados muestra la tabla de contingencia de las tres variables:

**Estrés * Edad * Sexo Crosstabulation**

Count

| Sexo | | | Edad | | | Total |
|---|---|---|---|---|---|---|
| | | | Adolescente | Jóven | Adulto | |
| Mujer | Estrés | No | 1 | 2 | | 3 |
| | | Si | | | 1 | 1 |
| | Total | | 1 | 2 | 1 | 4 |
| Hombre | Estrés | No | 1 | 1 | 2 | 4 |
| | | Si | 1 | 1 | | 2 |
| | Total | | 2 | 2 | 2 | 6 |

Lo que no permite el procedimiento Tablas de contingencia es realizar el análisis estadístico en más de dos variables cualitativas. Dicho análisis se lleva a cabo mediante los *modelos loglineales y modelos logit.*

## 8.7. EJERCICIOS PROPUESTOS

1. ¿La proporción de sujetos que reciben la terapia contra estados de ansiedad generalizada supera el valor 0,55 (= 0,05)?
2. ¿Puede afirmarse que la proporción de sujetos que recibe la terapia para reducir el estrés difiere de la que recibe la terapia para combatir la ansiedad generalizada? (= 0,05)
3. ¿Puede afirmarse que, al menos la mitad de los sujetos han recibido la terapia contra el estrés? (= 0,01)
4. Uno de los psicólogos del gabinete que está tratando con estos trabajadores con problemas de estrés opina que, en la población, el 60% de los pacientes son varones. Compruebe esta hipótesis con = 0,05.
5. Obtenga la tabla de distribución conjunta para los que tienen Fobia y no tienen Fobia en la variable Edad. ¿Puede afirmarse que existe relación entre ambas variables con = 0,01?
6. ¿Puede afirmarse que el porcentaje de varones que han recibido la terapia contra el estrés es diferente que el porcentaje de mujeres que la han recibido? (con = 0,01).
7. ¿Existe relación entre las variables Fobia y Recibir terapia contra el estrés? (con = 0,05).

CAPÍTULO

9

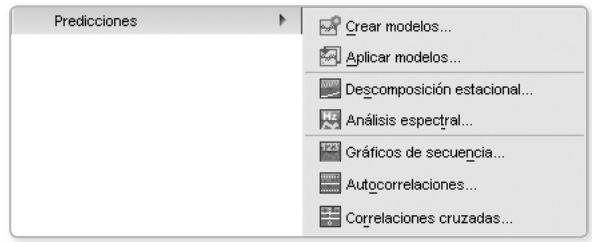
# SERIES TEMPORALES, SUPERVIVENCIA Y RESPUESTAS MÚLTIPLES

El menú Analizar es el más importante en SPSS, y se centra en los componentes estadísticos básicos de SPSS. Este es un capítulo práctico, donde se presentan numerosas técnicas estadísticas básicas y se discuten situaciones en las que son aplicables cada una de ellas, las hipótesis previas necesarias y cómo realizar e interpretar los resultados.

Las potentes técnicas analíticas de SPSS le ayudan a tomar las decisiones más adecuadas para hacer crecer su organización. La amplia gama de técnicas estadísticas permite ir más allá de las estadísticas descriptivas, y las filas y columnas. SPSS proporciona procedimientos que permiten resolver sus problemas en los negocios o la investigación.

## 9.1. SERIES TEMPORALES

Se define una **serie temporal** como una sucesión de observaciones cuantitativas de un fenómeno, ordenadas en el tiempo. Para la realización de un análisis clásico descriptivo de una serie temporal con SPSS, se tienen los siguientes procedimientos estadísticos:




### A. Suavizado exponencial

Suaviza componentes irregulares de datos de series temporales; para ello, hace uso de una variedad de modelos que incorporan diferentes supuestos acerca de la tendencia y estacionalidad. Seleccionando un modelo estacional y definiendo la periodicidad mediante Definir fechas, se puede seleccionar una variable que represente los factores estacionales.

Los modelos difieren en los componentes estacionales y de tendencia. Se dispone de los modelos Simple, Holt, Winters y Personalizado.

### B. Autorregresión

Estima un modelo de regresión lineal con errores autorregresivos de primer orden. Los métodos disponibles son Máxima verosimilitud exacta, Cochrane-Orcutt y Prais-Winsten.

Se debe desactivar Incluir constante en el modelo si no se desea estimar un término constante en la ecuación de regresión.

### C. Sequence Charts

Crea un gráfico de casos en secuencia. Este procedimiento requiere datos de serie temporal u otros, en los que se ordenen los casos con un criterio significativo. Representa, en dos dimensiones, los valores de cada una de las variables especificadas, frente al número de secuencia en el archivo de datos. Como se ha mencionado, si la variable es una serie temporal cuyos valores se le han asignado fechas (opción *Definir fechas*), la representación se realizará frente al instante de tiempo correspondiente.

### D. Arima

Estima modelos Arima (modelo autorregresivo integrado de media móvil) univariados estacionales y no estacionales (también conocidos como «modelos Box-Jenkins). Se pueden transformar las series antes de la estimación, mediante una de las opciones de la lista Transformar:

- **Ninguna**: No se lleva a cabo alguna transformación.
- **Log natural:** Transforma las series antes de la estimación con el logaritmo en base e.
- **Log base 10**: Transforma las series antes de la estimación con el logaritmo en base 10.

El grupo Modelo permite especificar los tres parámetros del modelo Arima: *Autorregresivo*, *Diferenciación* y *Media móvil*. Estos parámetros se conocen de forma más común como p, d y q, respectivamente. Asimismo, se pueden definir los parámetros estacionales correspondientes, si se tiene un modelo estacional y se ha definido la estacionalidad de los datos en Definir fechas. Además, si se ha definido la periodicidad (que se muestra en la parte inferior del cuadro), se pueden especificar los parámetros estacionales correspondientes, introduciendo los valores en los cuadros sp, sd y sq estacionales.

#### E. Descomposición estacional

Estima factores estacionales multiplicativos o aditivos para las series temporales. Precisa que esté definida la periodicidad mediante Definir fechas, y permite seleccionar el tipo de modelo utilizado para la descomposición estacional, que puede ser multiplicativo o aditivo.

Las opciones de Ponderación de la media móvil permiten especificar la manera de tratar las series, en el cálculo de medias móviles. Estas opciones solo están disponibles si la periodicidad de la serie es par (si es impar, todos los puntos son ponderados por igual):

- **Todos los puntos por igual.** Calcula las medias móviles con una amplitud igual a la periodicidad y todos los puntos ponderados por igual.
- **Puntos finales ponderados por 5.** Calcula las medias móviles con una amplitud igual a la periodicidad más uno y con los puntos finales ponderados por 0,5. Mediante Mostrar el listado por casos, se imprime un listado por casos que contiene un resumen en una línea en cada iteración, así como los estadísticos finales.

### 9.1.1. EJERCICIO PRÁCTICO 1

Se trabajará con el archivo de la encuesta del supermercado.

#### A. Serie Temporal

Se puede definir una serie temporal como una sucesión de observaciones cuantitativas de un fenómeno, ordenadas en el tiempo. Toda serie temporal, desde el enfoque del análisis clásico, puede ser descompuesta en cuatro componentes principales:

- **Ciclo**: Representa las oscilaciones de la serie que se producen con un período superior al año.
- **Tendencia**: Representa el movimiento general a largo plazo de una serie temporal.
- **Estacionalidad**: Representa oscilaciones que se producen con una periodicidad igual o inferior al año, y que se reproducen de manera reconocible cada año.
- **Residuo**: Movimiento que no muestra un carácter periódico reconocible.

El objeto del análisis de series temporales será la determinación de cada uno de estos componentes, con la finalidad de realizar predicciones o conocer, de manera descriptiva, el comportamiento de la serie temporal.

### B. Series temporales con SPSS

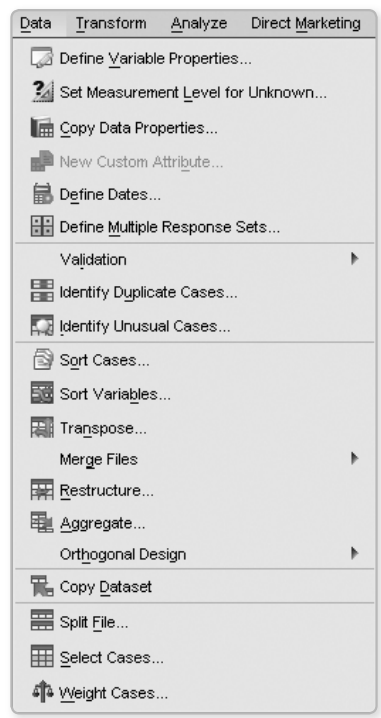
Si se quiere realizar un análisis clásico descriptivo de una serie temporal con SPSS, se deben seguir los siguientes pasos:

- Definición de fechas
- Representación gráfica de una serie temporal
- Distinción entre los esquemas aditivos y multiplicativos de una serie temporal
- Determinación de la tendencia
- Determinación de la estacionalidad

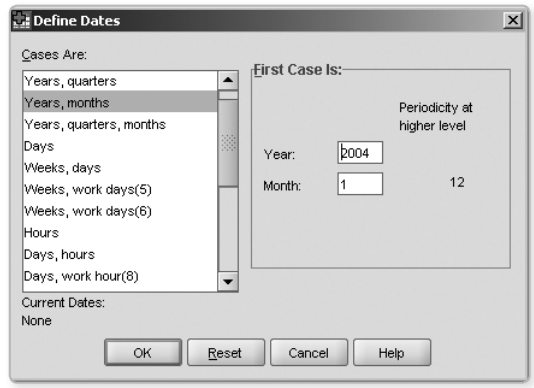
## 9.1.2. SERIES TEMPORALES CON SPSS

### A. Definición de fechas

Una serie temporal exige que las observaciones del fenómeno objeto de estudio estén ordenadas en el tiempo, debido a la definición de serie temporal. La realización de un análisis clásico descriptivo de una serie temporal, pasa por la necesidad de la definición de la fechas en SPSS. El procedimiento a seguir para la definición de las fechas de la serie temporal se definirá utilizando el comando **Definir fechas,** el cual se ejecuta al hacer un solo clic con el botón izquierdo del mouse cuando el puntero del mismo se encuentre sobre la opción del mismo nombre que aparece al desplegarse el menú **Datos** (Datos > Definir fechas).



El procedimiento Definir fechas de SPSS permite generar variables de fecha que se emplearán para establecer la periodicidad de una serie temporal, y para etiquetar los resultados del análisis de series temporales. Luego de la ejecución de este procedimiento, se obtendrá el cuadro de diálogo representado en la siguiente figura (cuadro de diálogo Definir fechas):




En el recuadro izquierdo (recuadro con encabezamiento Los casos son) de este cuadro de diálogo, se selecciona el formato de fecha deseado. Luego, se indica la fecha de la primera observación de la serie en el recuadro El primer caso es (First Case Is). Finalmente, debe hacer clic con el botón izquierdo del mouse, cuando el puntero se sitúe sobre Aceptar (OK). En el Visor de resultados, una anotación aparecerá en la que se indica el nombre y la etiqueta de las nuevas variables, creadas a consecuencia de la operación realizada (*year*¸ *month y date*). Observe el resultado:

The following new variables are being created:

| Name | Label |
|---|---|
| YEAR_ | YEAR, not periodic |
| MONTH_ | MONTH, period 12 |
| DATE_ | DATE. FORMAT: "MMM YYYY" |

En el Editor de datos se muestran los respectivos valores de estas variables, tal como se muestra en la siguiente tabla:

| | aseo | verduras | abarrote | bebidas | year_ | month | date_ |
|---|---|---|---|---|---|---|---|
| 1 | 2.00 | 17.00 | 34.50 | 6.00 | 2004 | 1 | JAN 2004 |
| 2 | 3.00 | 20.00 | 40.00 | 6.00 | 2004 | 2 | FEB 2004 |
| 3 | .30 | 10.00 | 31.60 | 6.00 | 2004 | 3 | MAR 2004 |
| 4 | 3.30 | 9.00 | 35.40 | 6.00 | 2004 | 4 | APR 2004 |
| 5 | 1.30 | 23.00 | 30.00 | 6.00 | 2004 | 5 | MAY 2004 |
| 6 | .40 | 13.00 | 32.90 | 6.00 | 2004 | 6 | JUN 2004 |
| 7 | 1.50 | 12.00 | 33.20 | 6.00 | 2004 | 7 | JUL 2004 |
| 8 | 4.50 | 19.00 | 33.10 | 6.00 | 2004 | 8 | AUG 2004 |
| 9 | 2.50 | 18.00 | 35.60 | 6.00 | 2004 | 9 | SEP 2004 |
| 10 | .30 | 24.00 | 33.00 | 6.00 | 2004 | 10 | OCT 2004 |

Las variables creadas en el Editor de datos al definir fechas (datos mensuales) son *year, month y date*. Si se realiza un análisis clásico descriptivo de una serie temporal, resultará incompleto si no se lleva a cabo la representación gráfica de la serie temporal, la cual persigue identificar patrones de comportamiento regulares que permitan plantear hipótesis sobre la estacionalidad y el ciclo, el esquema según el cual se combinan las componentes de la serie temporal, o la existencia de outliers o valores atípicos en la serie temporal, que adviertan de errores en la recolección de los datos o de los fenómenos imprevistos de un impacto excepcionalmente grande en los valores de la variable.

### B. Representación gráfica de una serie temporal

Para la representación gráfica de una serie temporal en SPSS se debe seleccionar el comando **Secuencia** de SPSS, el cual se ejecuta al hacer un solo clic con el botón izquierdo del mouse cuando el puntero del mismo se encuentre sobre la opción del mismo nombre que aparece al desplegarse el menú **Análisis** de la barra de menú de SPSS (**Serie de tiempo > Gráfico de secuencia**). La ejecución de este procedimiento requiere disponer de datos de serie temporal o bien de otro tipo de datos, en los cuales los casos aparezcan ordenados, según un criterio significativo. Una vez ejecutado el comando Secuencia, se muestra en la pantalla el cuadro de diálogo Gráficos de secuencia (ver fig.), donde se introducirá, en el recuadro Variables, la variable o variables cuantitativas a representar, y en el recuadro Etiquetas del eje de tiempo, una variable categórica o numérica, la cual se empleará para etiquetar los ejes temporales de la representación o representaciones gráficas.

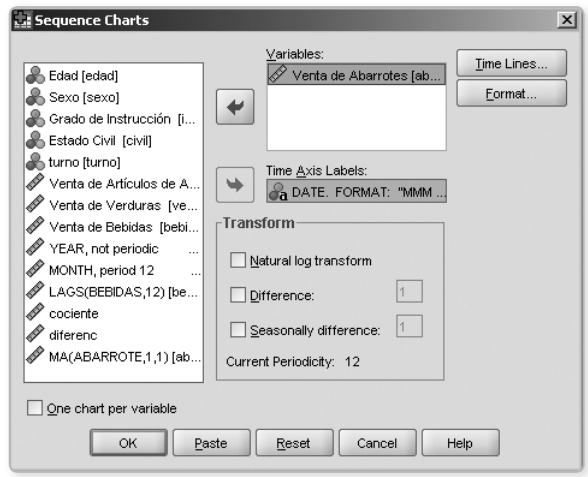



*Figura 9.1 Cuadro de diálogo Gráficos de secuencia*

Finalmente, sin modificar el resto de opciones que aparecen, por defecto en el cuadro de diálogo, se hará clic con el botón izquierdo del mouse cuando el puntero se encuentre sobre el botón Aceptar, para ejecutar el comando **Secuencia** que mostrará el gráfico de la serie temporal en el Visor de resultados.

### C. Distinción entre el esquema aditivo y esquema multiplicativo de una serie temporal

En lo referente al esquema según el cual se combinan los componentes de la serie temporal, pese a la existencia de métodos analíticos para la determinación del mismo, se va a optar por el método gráfico, el cual consiste en identificar el esquema de la serie a través de la observación de la representación gráfica de la misma. De esta forma, una serie posee un esquema aditivo si la variabilidad de la misma permanece constante a lo largo del tiempo, tal y como se muestra en la figura.

Gráfica de una serie aditiva

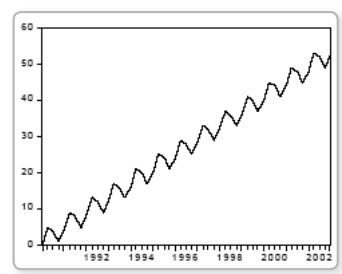



*Figura 9.2 Gráfica de una serie aditiva*

Se dice que la serie posee un esquema multiplicativo cuando esta variabilidad aumenta con el tiempo, tal como se muestra en la figura:

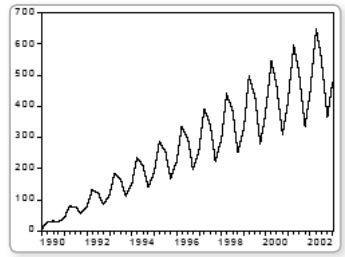



*Figura 9.3 Gráfica de una serie multiplicativa*

En la práctica, para diferenciar los modelos de series aditivos y multiplicativos, se emplean las técnicas del método analítico de las diferencias y cocientes estacionales, y el método gráfico.

**a. Método analítico de las diferencias y cocientes estacionales**

Esta técnica permite determinar el tipo de esquema según el cual se combinan los componentes del análisis clásico de series temporales, mediante la comparación de los coeficientes de variación de Pearson, de las series de las diferencias ($d = y_t - y_{tp}$) y de los cocientes [$C = y_t / y_{tp}$] estacionales.

$$CV_d = \frac{S_d}{\overline{Y}_d} \quad o \quad CV_d = \frac{S_d}{\overline{Y}_d}$$

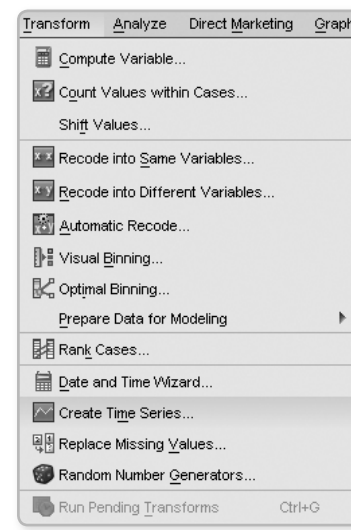
De esta forma, el esquema será aditivo o multiplicativo, de acuerdo con la siguiente regla de decisión:

- Si $CV_d > CV_c$, entonces el esquema es multiplicativo
- Si $CV_c > CV_d$, entonces el esquema es aditivo

Los pasos a seguir en SPSS son los que se indican a continuación:

**Paso 1:** Seleccione como valor de *p* aquel que represente el periodo de oscilaciones más importante de la serie (12 si la serie es mensual, 4 si es trimestral, 3 si es cuatrimestral, etc.).

**Paso 2:** Cree la variable de retardos $y_{t-p}$. Para su obtención, habrá que hacer un solo clic con el botón izquierdo del mouse sobre la opción **Crear serie temporal,** que se despliega cuando su puntero se encuentra situado sobre el menú Transformar de la barra de menú de SPSS (Transformar > Crear serie temporal). Seleccione la función Retardo y elija como orden el valor de *p* fijado previamente en el paso 1.



A continuación, añada la variable y = Bebida a la ventana Nuevas variables. Tras realizar esta operación, se activa el cuadro de diálogo que se representa en la siguiente figura:

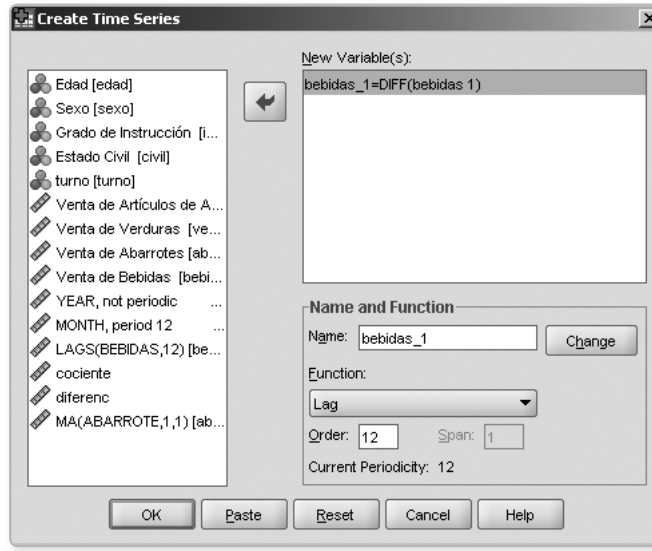



*Figura 9.4 Obtención de la serie de los retardos*

**Paso 3:** Cree la serie de las diferencias y de los cocientes. Para su obtención, debe hacer un solo clic con el botón izquierdo del mouse sobre la opción **Calcular,** que se despliega cuando su puntero se encuentra situado sobre el menú **Transformar,** de la barra de menú de SPSS (Transformar > Calcular).

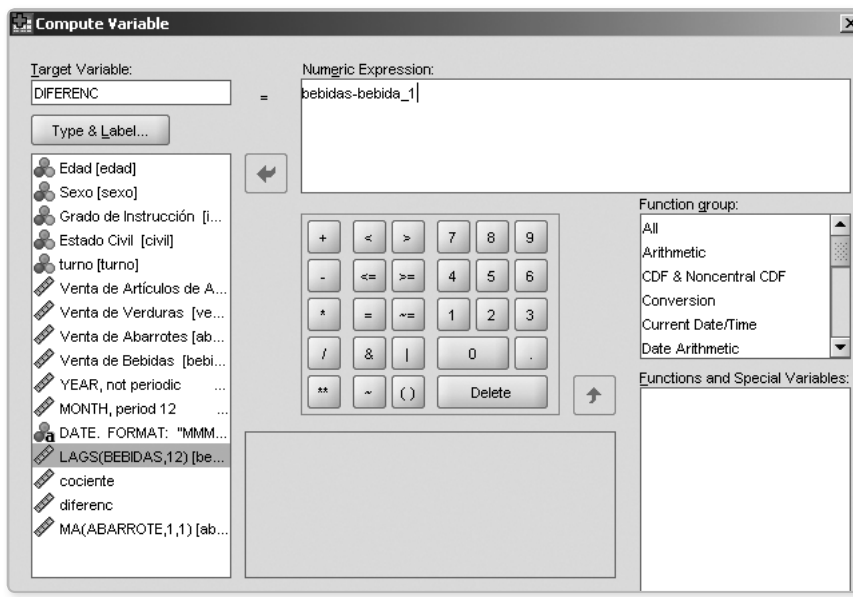



*Figura 9.5 Obtención de la serie de diferencias*

A continuación, se crean dos variables nuevas a partir de las variables ya existentes (*y* = Bebida) *y* = Bebida_1 llamada «diferencia», la cual se obtiene como diferencia de las variables $y_t$ - $y_{tp}$ y «cociente», obtenida con el cociente $y_t$ / $y_{tp}$.

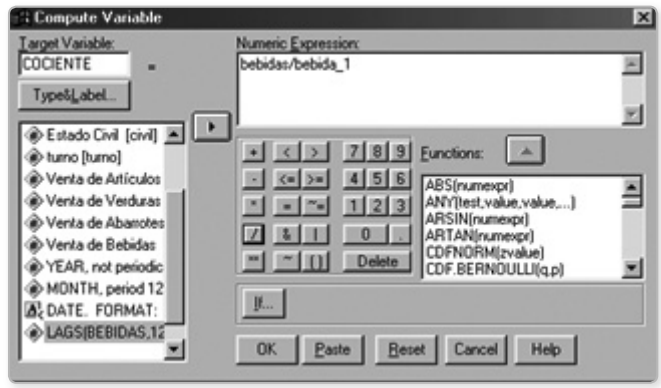



*Figura 9.6 Obtención de la serie de los cocientes*

**Paso 4:** Calcule el coeficiente de variación de Pearson, a partir de la media y desviación típica de las variables *Diferencia* y *Cociente* ya obtenidas en el paso 3. Para su obtención, haga un solo clic con el botón izquierdo del mouse sobre la opción **Estadísticos descriptivos,** que se despliega cuando su puntero se encuentra situado sobre el menú **Analizar** de la barra de menú de SPSS (Analizar > Estadísticos descriptivos > Descriptivos). Tras realizar esta operación, se activa el cuadro de diálogo que se representa en la siguiente figura:

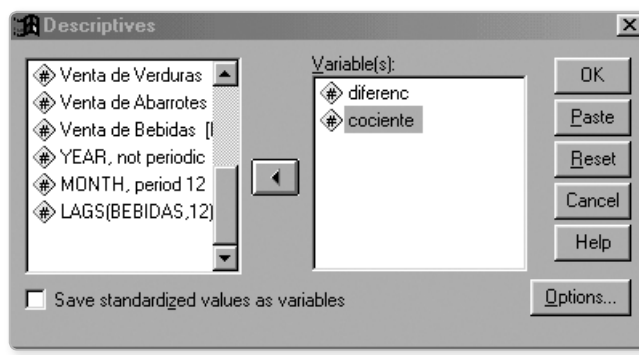



*Figura 9.7 Obtención de los estadísticos Media y Desviación típica*

Finalmente, seleccione con el botón Opciones los estadísticos *media* y *desviación típica* para las variables *Diferencia* y *Cociente*.

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| DIFERENC | 18 | -5.00 | 2.00 | -1.8333 | 2.6624 |
| COCIENTE | 18 | .17 | 3.00 | .9907 | .9540 |
| Valid N (listwise) | 18 | | | | |

**b. Método gráfico de la media y de la desviación típica**

Esta técnica consiste en representar gráficamente la media aritmética y la desviación típica de grupos de observaciones de la serie temporal. Si en el gráfico se observa (no se observa) una clara tendencia creciente o decreciente, entonces el esquema según el cual se combinan las componentes de la serie será multiplicativo (aditivo).

## D. Determinación de la tendencia de una serie temporal

Para la determinación de la tendencia de la serie temporal se pueden usar alternativamente dos métodos: El Método del Ajuste Analítico y el Método de la Media Móvil.

**a. Método de ajuste analítico**

La idea fundamental de este método consiste en ajustar una curva a las medias aritméticas anuales de las observaciones que componen la serie temporal, por el procedimiento de los mínimos cuadrados ordinarios (MCO). Para la aplicación de esta metodología en SPSS, habrá que seguir los siguientes pasos:

**Paso 1:** Con el objeto de eliminar el efecto producido por la componente estacional que distorsiona la tendencia, se hace necesario tomar medias anuales, para lo que se tendrá que recurrir al comando **Agregar**.

Para ejecutar el comando **Agregar**, habrá que hacer un solo clic con el botón izquierdo del mouse, cuando el puntero del mismo se encuentra sobre el menú Datos de la barra de menú (Datos > Agregar). Tras ejecutar el comando Agregar, se muestra el cuadro de diálogo representado en la figura.

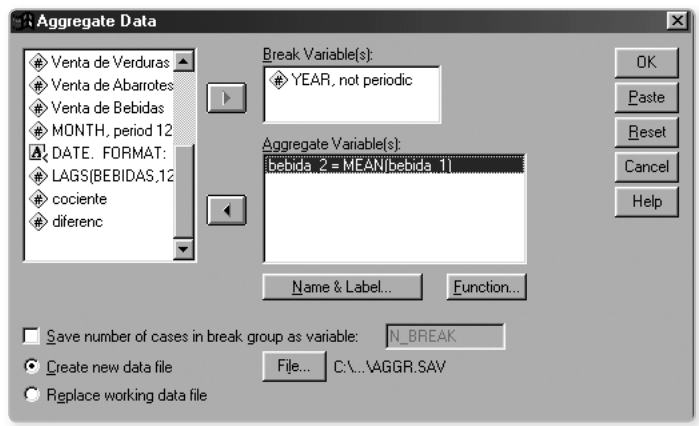



*Figura 9.8 Cuadro de diálogo Agregar datos*

En el cuadro de diálogo Agregar datos, se introducirá la(s) variable(s) que va(n) a definir los grupos a agregar en el recuadro Variables de segmentación, mientras que en el recuadro Agregar variables, se introducirá la(s) variable(s) que va(n) a ser agregadas indicando nombre, etiquetas y función de agregación (en este caso, la media la cual aparece seleccionada por defecto) de las mismas cuando estén seleccionadas, haciendo clic con el botón izquierdo del mouse con su puntero sobre el botón Nombre *y* Etiqueta y función, respectivamente.

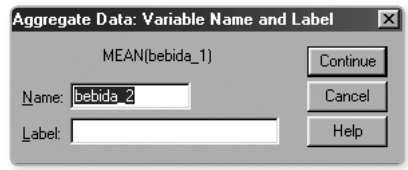



*Figura 9.9 Nombre y etiquetas de variables del cuadro de diálogo Agregar datos*

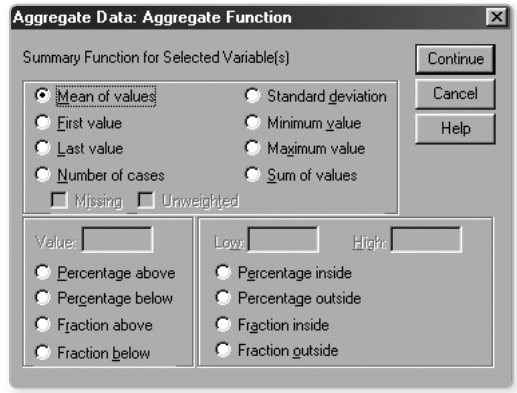
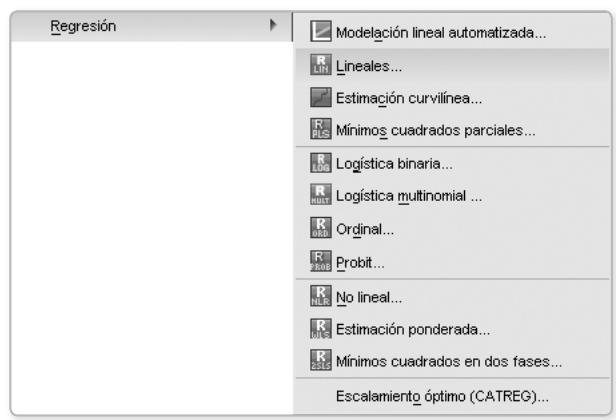



*Figura 9.10 Función de agregación del cuadro de diálogo Agregar datos*

Una vez definido el Nombre y etiqueta de la nueva variable creada y la función de agregación de la misma, se ejecutará el comando haciendo un clic con el botón izquierdo del mouse cuando su puntero se encuentre situado sobre el botón Aceptar del cuadro de diálogo **Agregar**. El resultado será guardado en la ubicación del disco duro que se indique antes de hacer clic en Aceptar, haciendo clic con el botón izquierdo del mouse cuando su puntero se encuentre situado sobre el botón **Archivo** de este cuadro de diálogo. Tras activar este botón, se mostrará el cuadro de diálogo en la pantalla (ver fig.), donde se indicará la ubicación en la que se desea guardar el resultado de la operación y el nombre que recibirá el archivo que recoge esta operación. Por ejemplo, al nuevo archivo se le puede denominar *Nombre Supermerdo_agregado.sav*.

**Paso 2:** Luego de abrir el nuevo archivo creado (*Supermerdo_agregado*), se procederá a aplicar un análisis de la regresión, donde la variable dependiente sea la variable que recoge las medias anuales (Bebidas), y la variable independiente sea la que representa el tiempo (year _ 1). Con la opción **Regresión lineal** que se encuentra situada sobre la opción **Analizar** del menú SPSS (Analizar regresión lineal).



Una vez realizada esta operación, se activa el cuadro de diálogo siguiente:




*Figura 9.11 Cuadro de diálogo Regresión lineal*

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .668[a] | .446 | .426 | 1.3354 |

a. Predictors: (Constant), YEAR, not periodic

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 40.238 | 1 | 40.238 | 22.566 | .000[a] |
| | Residual | 49.929 | 28 | 1.783 | | |
| | Total | 90.167 | 29 | | | |

a. Predictors: (Constant), YEAR, not periodic

b. Dependent Variable: Venta de Bebidas

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 3106.833 | 653.149 | | 4.757 | .000 |
| | YEAR, not periodic | -1.548 | .326 | -.668 | -4.750 | .000 |

a. Dependent Variable: Venta de Bebidas

**Paso 3:** Represente gráficamente e interprete los valores pronosticados en el análisis de regresión (la tendencia) y los valores agregados anuales de la serie, utilizando el comando gráfico Secuencia de SPSS.

Analizar > Regresión lineal > Guardar valores pronosticados no tipificados: Gráfico > Secuencia

Los valores pronosticados no tipificados de este análisis de regresión son los valores de la tendencia de cada año.

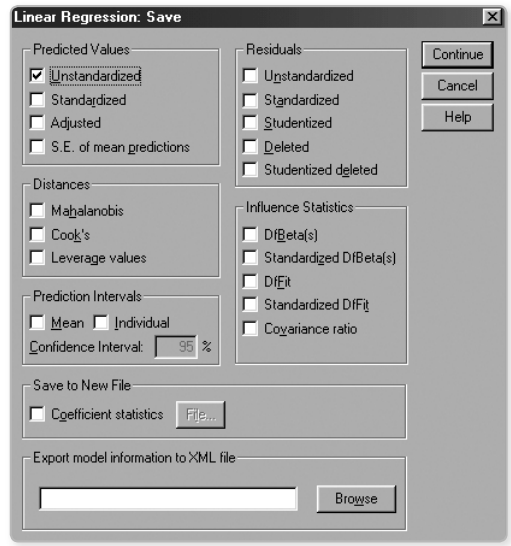



*Figura 9.12 Cuadro de diálogo Guardar nuevas variables*

**b. Método de la media móvil para determinar la tendencia**

Previamente al método de la media móvil, hay que utilizar el comando **Definir fechas** para definir el tiempo de una serie temporal. Este método consiste en determinar la tendencia de la serie por medio del uso de medias móviles de orden "*p*". Para la aplicación de esta metodología, debe dar los siguientes pasos:

**Paso 1:** Elija un valor de *p* que represente el periodo de oscilaciones más importante que posee la serie (normalmente, es el número de observaciones existentes dentro de cada año en ausencia de ciclo).

**Paso 2:** Calcule las medias móviles centradas mediante el comando **Crear serie temporal** de SPSS. Para ejecutar este comando, debe hacer un solo clic con el botón izquierdo del mouse sobre la opción del mismo nombre, que se despliega cuando su puntero se encuentra situado sobre el menú Transformar de la barra de menú de SPSS (Transformar > Crear serie temporal). Seguidamente, se activa el cuadro de diálogo siguiente:

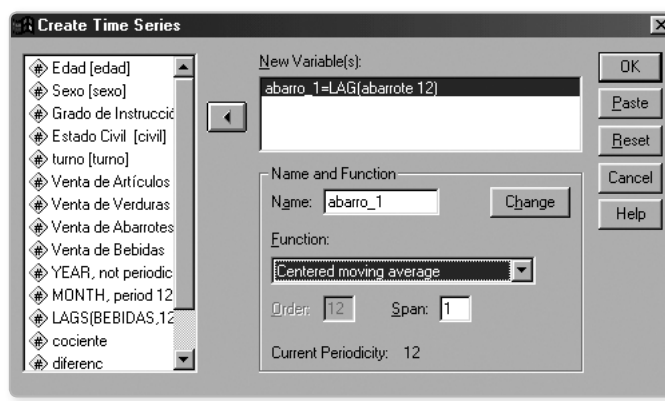



*Figura 9.13 Cuadro de diálogo Crear serie temporal*

En el cuadro de diálogo Crear serie temporal, se introducirá en el recuadro Nuevas variables, la variable cuya tendencia se quiere determinar, y en el recuadro Nombre y función se indicará el nombre de la nueva variable a crear, que recogerá la tendencia de la serie (MMC), la función de la serie que recoge la tendencia de la serie (media móvil centrada) y la amplitud (p = 12 para series mensuales, 4 para series trimestrales, etc.). Una vez completado el cuadro de diálogo con las especificaciones indicadas, se ejecutará el comando haciendo clic con el botón izquierdo del mouse, cuando el puntero se encuentre situado sobre el botón Aceptar. Como resultado, se creará una nueva variable con el nombre que se haya indicado, que recogerá la tendencia de la serie temporal.

**Paso 3:** Represente gráficamente los valores de la tendencia recogidos en la variable MMC y los valores que toma la serie temporal en los distintos momentos del tiempo, utilizando el comando **Secuencia** de SPSS (Gráfico > Secuencia).

## E. Determinación de la componente estacional de una serie temporal

De todos los métodos estudiados para la descomposición estacional de una serie temporal, el único que va a ser expuesto para su resolución con SPSS va a ser el método de la media móvil.

### a. Método de la media móvil para determinar la estacionalidad

Este método requiere, en primer lugar, obtener la componente extra estacional de la serie (ciclo + tendencia), utilizando el método de las medias móviles para la tendencia estudiado en el apartado anterior, y en segundo lugar, calcular los índices de variación estacional (IVE) según el esquema, de acuerdo con el que combinan las componentes de la serie. La aplicación de este método con SPSS implica dar los siguientes pasos:

**Paso 1:** Ejecute el comando **Descomposición estacional** de SPSS haciendo clic con el botón izquierdo del mouse en la opción del mismo nombre, que se despliega al desplazar el puntero del mouse sobre la opción **Series temporales** de la barra de menú **Analizar** de SPSS (Analizar > Series temporales > Descomposición estacional).

«El cálculo de los IVE en SPSS se lleva a cabo utilizando el promedio medial definido como la media de todos los valores, menos el máximo y el mínimo».

Tras realizar la operación descrita en el párrafo anterior, se mostrará en la pantalla el siguiente cuadro de diálogo:

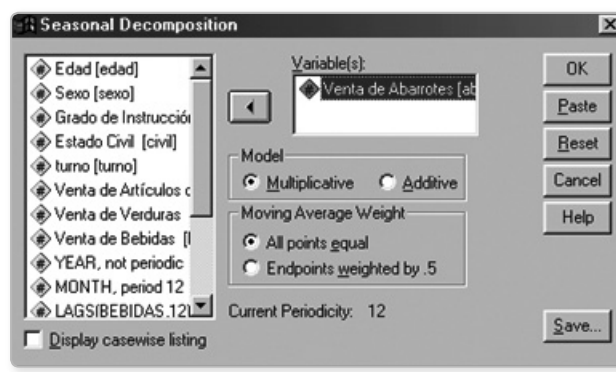



*Figura 9.14 Cuadro de diálogo Descomposición estacional*

En el cuadro de diálogo Descomposición estacional, se introducirán en el recuadro Variables la(s) variable(s) a las que se les va a realizar la descomposición estacional, y se indicará en la opción Modelo, el esquema que sigue la serie (multiplicativo o aditivo).

**NOTA**

*Si alguna de las variables introducidas, en la casilla Variables, tiene valores perdidos, el procedimiento de Descomposición estacional no se ejecuta.*

A continuación, en el recuadro **Ponderación de la media móvil**, se seleccionará una de las dos opciones siguientes:

- **Todos los puntos por igual.** Cuando el número de subdivisiones hechas dentro del año (*p*), sea un número impar.
- **Puntos finales ponderados por 5.** En caso de que *p* sea un número par.

Finalmente, pulsando el botón **Guardar** del cuadro de diálogo representado anteriormente, se indicará cómo se van a tratar las nuevas variables creadas como consecuencia de la descomposición estacional de la serie, pudiéndose optar por:

- **Añadir al archivo**. Añade las nuevas variables creadas mediante este procedimiento al archivo activo.
- **Reemplazar las existentes.** Guarda solo las variables del procedimiento actual Si en el archivo existen ya estas variables, se sustituirían.
- **No crear**. Indica que no se deberá guardar alguna variable.

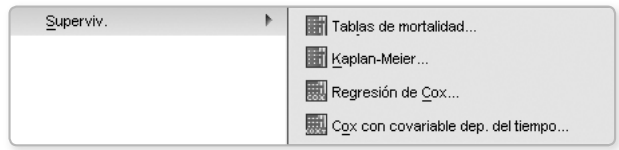
**Paso 2:** Interpretación de los resultados de la descomposición estacional. Como consecuencia de la descomposición estacional se crean cuatro nuevas variables que se añaden al archivo (en el caso de que se haya seleccionado la opción Añadir al archivo) en el cuadro de diálogo que aparece al pulsar el botón **Guardar** del cuadro de diálogo que aparece al pulsar el botón Guardar del cuadro de diálogo Descomposición estacional tal y como se muestra a continuación en la figura. El contenido de cada una de estas nuevas variables es el que denota a continuación:

- **ERR**. Contiene la componente residual o errática de la serie. Es posible su cálculo utilizando el resto de las componentes si se tiene en cuenta el esquema de composición de la misma.
- **SAS**. Recoge los valores de la serie desestacionalizada y se calcula por la diferencia entre los valores de estacionalidad en el caso de que el esquema sea aditivo y por el cociente en el caso de que el esquema sea multiplicativo.
- **SAF**. Recoge los factores de estacionalidad de la serie los cuales se repiten cada 3, 4, 12... veces, según sea la serie cuatrimestral, trimestral, anual.

Su interpretación depende si, en la descomposición estacional, se ha considerado un esquema aditivo o multiplicativo:

- **Aditivo.** Cuando toma el valor igual a 0, no existe estacionalidad en el periodo concreto en el que lo toma; cuando este es positivo en un periodo concreto, el valor de la variable toma valores superiores a los de la media en ese periodo; mientras que cuando es negativo, ocurre lo contrario.
- **Multiplicativo.** Cuando toma el valor igual a 1, no existe estacionalidad en el periodo concreto en el que lo toma; cuando este es mayor que 1 en un periodo concreto, el valor de la variable toma valores superiores a los de la media en ese periodo; mientras que cuando es menor que 1 ocurre lo contrario.
- **STC**. Recoge la componente tendencia –ciclo resultante de aplicar un análisis de la tendencia utilizando el método de la media móvil. El comando Descomposición estacional; permite, por lo tanto, la determinación de la tendencia, a través del método de la media móvil cuando el valor de p es igual al número de subdivisiones que se realicen dentro del año.

## 9.2. SUPERVIVENCIA




### A. Tablas de mortalidad

Existen muchas situaciones en las se desea examinar la distribución de un periodo entre dos eventos, como la duración del empleo (tiempo transcurrido entre el contrato y el abandono de la empresa). Sin embargo, este tipo de datos suelen incluir algunos casos para los que no se registra el segundo evento (censurados); por ejemplo, la gente que todavía trabaja en la empresa al final del estudio.

Las razones para que no se verifique el segundo evento pueden ser muy variadas: el evento simplemente no tiene lugar antes de que finalice el estudio, el investigador puede haber perdido el seguimiento de su estado en algún momento anterior a que finalice el estudio, no pueden continuar por razones ajenas al estudio (cuando un empleado cae enfermo y se acoge a una baja laboral). Estos casos se conocen globalmente como «casos censurados» y hacen que el uso de técnicas tradicionales como las pruebas t o la regresión lineal sean inapropiadas para este tipo de estudio. Existe una técnica estadística útil para este tipo de datos llamada «tabla de mortalidad de seguimiento». La idea básica de la **tabla de mortalidad** es subdividir el periodo de observación en intervalos de tiempo más pequeños. En cada intervalo, se utiliza toda la gente que se ha observado como mínimo durante ese periodo de tiempo, para calcular la probabilidad de que un evento terminal tenga lugar dentro de ese intervalo.

Las probabilidades estimadas para cada intervalo se emplean para calcular la probabilidad global de que el evento tenga lugar en diferentes puntos temporales.

### B. Kaplan-Meier

Es un método de estimación de modelos hasta el evento, en presencia de casos censurados. Se basa en la estimación de las probabilidades condicionales en cada punto temporal, cuando tiene lugar un evento, y en tomar el límite del producto de esas probabilidades para estimar la tasa de supervivencia en cada punto temporal.

### C. Regresión de Cox

Del mismo modo que los anteriores, la **regresión de Cox** es un método para crear modelos sobre datos de tiempos de espera hasta un evento con casos censurados presentes. Sin embargo, la regresión de Cox permite incluir variables predictoras (covariables) en los modelos.

Por ejemplo, permite construir un modelo de la duración en el empleo como función del nivel educativo y de la categoría laboral. Gestiona los casos censurados correctamente y proporciona las estimaciones de los coeficientes para cada una de las covariables, permitiendo evaluar el impacto de múltiples covariables en el mismo modelo. Además, es posible utilizar este método para examinar el efecto de covariables continuas.

**D. Cox con covariable dep. del tiempo**

Existen ciertas situaciones en las que interesa calcular un modelo de regresión de Cox, pero no se cumple el supuesto de tasas de impacto proporcionales; es decir, las tasas de impacto cambian con el tiempo: los valores de una (o de varias) de las covariables son diferentes en los distintos puntos del tiempo. En esos casos, es necesario utilizar un modelo de regresión de Cox extendido que permita especificar covariables dependientes del tiempo (T_COV_).

Con el fin de analizar dicho modelo, se debe definir primero una covariable dependiente del tiempo. Para facilitar esta tarea, se cuenta con una variable del sistema llamada T_ que representa el tiempo. Se utiliza para definir covariables dependientes del tiempo empleando dos métodos generales, a saber:

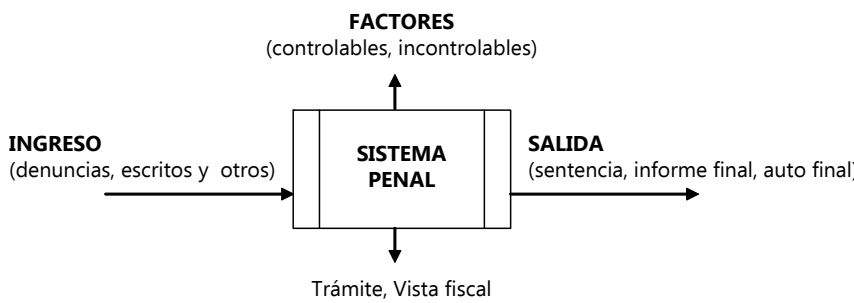
- Para contrastar el supuesto de tasas de impacto proporcionales con respecto a una covariable particular o para estimar un modelo de regresión de Cox extendido que permita impactos no proporcionales. Se define la **covariable dependiente del tiempo** como una función de la variable de tiempo T_ y la covariable en cuestión.
- Algunas variables pueden tener valores distintos en periodos diferentes del tiempo, pero no están sistemáticamente relacionadas con el tiempo. En tales casos, es necesario definir una covariable dependiente del tiempo segmentada, lo cual puede llevarse a cabo a partir de un conjunto de medidas, usando las expresiones lógicas (1: verdadero, 0: falso).

Para las covariables dependientes del tiempo segmentadas, los casos en los que falte cualquiera de los valores quedarán eliminados del análisis. Por tanto, todos los casos deben tener valores para todos los puntos del tiempo medidos en la covariable, incluso para los puntos del tiempo posteriores a la eliminación del caso del conjunto bajo riesgo (ya sea, por evento o censura).

### 9.2.1. EJERCICIO PRÁCTICO 1

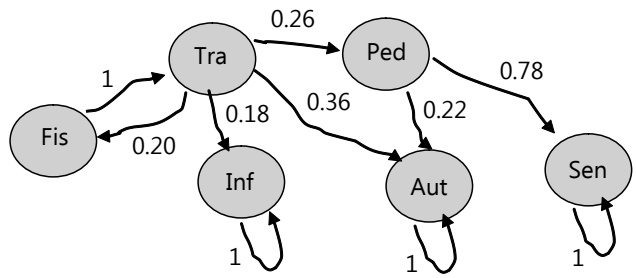
Se tiene el siguiente sistema:

SISTEMA PENAL**.** El proceso se inicia con el ingreso de una denuncia, escrito u otro tipo de servicio; dicho ingreso es recibido por la Mesa de Partes Única de los Juzgados Penales, que designa el Juzgado que tendrá competencia del proceso. Si el ingreso es una denuncia, el reparto se hace en forma aleatoria entre las dependencias existentes. Luego de ser recibido por el Juzgado, este dispondrá el trámite procesal que debe seguir según las normas del Código de Procedimientos Penales y de acuerdo con el tipo de proceso (sumario, ordinario y especial). En cada dependencia penal, el recorrido o estados del proceso por donde deben pasar los expedientes o procesos son los mismos. En el diagrama que se muestra a continuación, se aprecia la estructura del sistema de información penal.



Entonces, se ilustra la situación utilizando la teoría de grafos.

**ESTADOS DEL SISTEMA (RECORRIDO)**




Para el presente estudio, se toma como 1: Muerto cuando el proceso termina y puede ser sentenciado, informe final, auto final y 0: Censurada cuando el proceso entra a reserva.

Tipo de código (Dosis):

- 0: Nuevo Código procesal Penal
- 1: Código de Procedimientos Penales (antiguo)

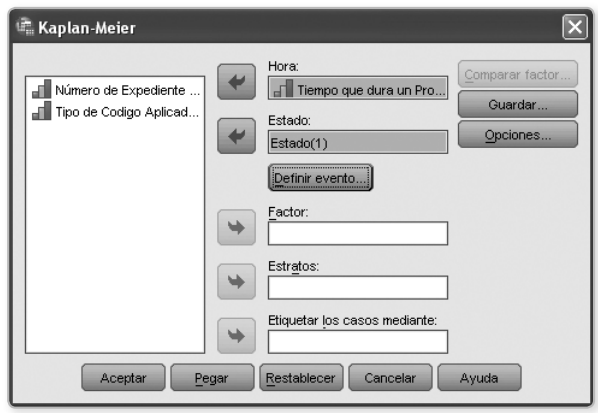
| N° de expediente | Duración del proceso | Estado | Tipo de código |
|---|---|---|---|
| 1 | 9 | 1 | 1 |
| 2 | 12 | 1 | 1 |
| 3 | 12 | 0 | 1 |
| 4 | 13 | 0 | 1 |
| 5 | 14 | 0 | 1 |
| 6 | 16 | 1 | 1 |
| 7 | 18 | 0 | 1 |
| 8 | 19 | 1 | 1 |
| 9 | 23 | 0 | 1 |
| 10 | 24 | 0 | 1 |
| 11 | 25 | 0 | 1 |
| 12 | 26 | 0 | 1 |
| 13 | 27 | 1 | 1 |
| 14 | 29 | 0 | 1 |
| 15 | 30 | 0 | 1 |
| 16 | 32 | 0 | 1 |
| 17 | 33 | 0 | 1 |
| 18 | 33 | 0 | 1 |
| 19 | 35 | 1 | 1 |

| | | | |
|---|---|---|---|
| 20 | 35 | 1 | 1 |
| 21 | 35 | 0 | 1 |
| 22 | 35 | 0 | 1 |
| 23 | 35 | 0 | 1 |
| 24 | 36 | 1 | 1 |
| 25 | 38 | 0 | 1 |
| 26 | 51 | 0 | 1 |
| 27 | 54 | 0 | 1 |
| 28 | 57 | 1 | 1 |
| 29 | 60 | 0 | 1 |
| 30 | 67 | 1 | 1 |
| 31 | 70 | 1 | 1 |
| 32 | 87 | 0 | 1 |
| 33 | 89 | 0 | 1 |
| 34 | 98 | 0 | 1 |
| 35 | 120 | 0 | 1 |
| 36 | 7 | 1 | 0 |
| 37 | 9 | 1 | 0 |
| 38 | 12 | 1 | 0 |
| 39 | 12 | 1 | 0 |
| 40 | 19 | 1 | 0 |
| 41 | 23 | 1 | 0 |
| 42 | 24 | 1 | 0 |
| 43 | 24 | 1 | 0 |
| 44 | 24 | 1 | 0 |
| 45 | 24 | 1 | 0 |
| 46 | 29 | 0 | 0 |
| 47 | 34 | 1 | 0 |
| 48 | 41 | 1 | 0 |
| 49 | 54 | 1 | 0 |
| 50 | 72 | 0 | 0 |
| 51 | 78 | 1 | 0 |
| 52 | 80 | 0 | 0 |
| 53 | 83 | 0 | 0 |
| 54 | 92 | 0 | 0 |
| 55 | 139 | 0 | 0 |
| 56 | 139 | 0 | 0 |

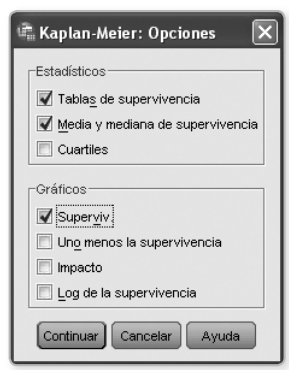
## Resultados

### A. Curva de supervivencia por el método de Kaplan-Meier

Para llevar a cabo el procedimiento, en la opción Analizar, elija Supervivencia y luego Kaplan-Meier. El estado al final del seguimiento es una variable categórica dicotómica, soliendo asignarse la categoría 1 a quienes tuvieron el evento evaluado (muerte) y la categoría 0 a los llamados casos censurados, ya sea porque se perdieron durante el seguimiento o porque no habían tenido el evento final al culminar el estudio.




Luego:




### B. El resultado que se obtiene es la siguiente salida de SPSS:

Seguidamente, el programa SPSS informa del número total de sujetos evaluados, del número de censurados y el número de desenlaces.

**Resumen del procesamiento de los casos**

| | | Censurado | |
|---|---|---|---|
| Nº total | Nº de eventos | Nº | Porcentaje |
| 56 | 25 | 31 | 55,4% |

En la tabla, la segunda columna de la izquierda (Tiempo) es el tiempo durante el cual los diferentes individuos están en seguimiento, ordenados por tiempos crecientes. La tercera columna (Estado) indica si se ha producido el efecto o desenlace evaluado o no; en este caso, etiquetados como «muerte» (proceso concluido). La siguiente columna es la supervivencia acumulada o proporción de casos para los que no ha tenido lugar el evento en cada tiempo. La siguiente columna es el error estándar correspondiente a la estimación puntual de Kaplan-Meier en cada tiempo. La sexta columna son los desenlaces acumulados; esto es, los que han muerto hasta ese tiempo. La última columna es el número de expedientes que quedan en cada momento, sin que haya ocurrido en ellos el evento final o desenlace evaluado, y representan los expedientes que *están a riesgo en el siguiente periodo*.

**Tabla de supervivencia**

| | Tiempo | Estado | Proporción acumulada que sobrevive hasta el momento | | Nº de eventos acumulados | Nº de casos que permanecen |
|---|---|---|---|---|---|---|
| | | | Estimación | Error típico | | |
| 1 | 7,000 | Concluido | ,982 | ,018 | 1 | 55 |
| 2 | 9,000 | Concluido | . | . | 2 | 54 |
| 3 | 9,000 | Concluido | ,946 | ,030 | 3 | 53 |
| 4 | 12,000 | Concluido | . | . | 4 | 52 |
| 5 | 12,000 | Concluido | . | . | 5 | 51 |
| 6 | 12,000 | Concluido | ,893 | ,041 | 6 | 50 |
| 7 | 12,000 | Reserva | . | . | 6 | 49 |
| 8 | 13,000 | Reserva | . | . | 6 | 48 |
| 9 | 14,000 | Reserva | . | . | 6 | 47 |
| 10 | 16,000 | Concluido | ,874 | ,045 | 7 | 46 |
| 11 | 18,000 | Reserva | . | . | 7 | 45 |
| 12 | 19,000 | Concluido | . | . | 8 | 44 |
| 13 | 19,000 | Concluido | ,835 | ,050 | 9 | 43 |
| 14 | 23,000 | Concluido | ,816 | ,053 | 10 | 42 |
| 15 | 23,000 | Reserva | . | . | 10 | 41 |
| 16 | 24,000 | Concluido | . | . | 11 | 40 |
| 17 | 24,000 | Concluido | . | . | 12 | 39 |
| 18 | 24,000 | Concluido | . | . | 13 | 38 |
| 19 | 24,000 | Concluido | ,736 | ,061 | 14 | 37 |
| 20 | 24,000 | Reserva | . | . | 14 | 36 |
| 21 | 25,000 | Reserva | . | . | 14 | 35 |
| 22 | 26,000 | Reserva | . | . | 14 | 34 |
| 23 | 27,000 | Concluido | ,714 | ,063 | 15 | 33 |
| 24 | 29,000 | Reserva | . | . | 15 | 32 |
| 25 | 29,000 | Reserva | . | . | 15 | 31 |
| 26 | 30,000 | Reserva | . | . | 15 | 30 |
| 27 | 32,000 | Reserva | . | . | 15 | 29 |
| 28 | 33,000 | Reserva | . | . | 15 | 28 |
| 29 | 33,000 | Reserva | . | . | 15 | 27 |
| 30 | 34,000 | Concluido | ,688 | ,066 | 16 | 26 |
| 31 | 35,000 | Concluido | . | . | 17 | 25 |
| 32 | 35,000 | Concluido | ,635 | ,071 | 18 | 24 |
| 33 | 35,000 | Reserva | . | . | 18 | 23 |
| 34 | 35,000 | Reserva | . | . | 18 | 22 |
| 35 | 35,000 | Reserva | . | . | 18 | 21 |
| 36 | 36,000 | Concluido | ,605 | ,073 | 19 | 20 |
| 37 | 38,000 | Reserva | . | . | 19 | 19 |
| 38 | 41,000 | Concluido | ,573 | ,076 | 20 | 18 |
| 39 | 51,000 | Reserva | . | . | 20 | 17 |
| 40 | 54,000 | Concluido | ,539 | ,079 | 21 | 16 |
| 41 | 54,000 | Reserva | . | . | 21 | 15 |
| 42 | 57,000 | Concluido | ,503 | ,081 | 22 | 14 |
| 43 | 60,000 | Reserva | . | . | 22 | 13 |
| 44 | 67,000 | Concluido | ,465 | ,084 | 23 | 12 |
| 45 | 70,000 | Concluido | ,426 | ,085 | 24 | 11 |
| 46 | 72,000 | Reserva | . | . | 24 | 10 |
| 47 | 78,000 | Concluido | ,383 | ,087 | 25 | 9 |
| 48 | 80,000 | Reserva | . | . | 25 | 8 |
| 49 | 83,000 | Reserva | . | . | 25 | 7 |
| 50 | 87,000 | Reserva | . | . | 25 | 6 |
| 51 | 89,000 | Reserva | . | . | 25 | 5 |
| 52 | 92,000 | Reserva | . | . | 25 | 4 |
| 53 | 98,000 | Reserva | . | . | 25 | 3 |
| 54 | 120,000 | Reserva | . | . | 25 | 2 |
| 55 | 139,000 | Reserva | . | . | 25 | 1 |
| 56 | 139,000 | Reserva | . | . | 25 | 0 |

Si ha marcado, le dará los valores del tiempo medio de supervivencia y de la mediana de supervivencia (el tiempo en el que el 50% de los expedientes siguen sin haber desarrollado el evento), con sus correspondientes errores estándares y los intervalos de confianza al 95%, señalándose que el estudio está limitado por el periodo máximo de seguimiento (en este ejemplo: 139 meses).

Medias y medianas del tiempo de supervivencia

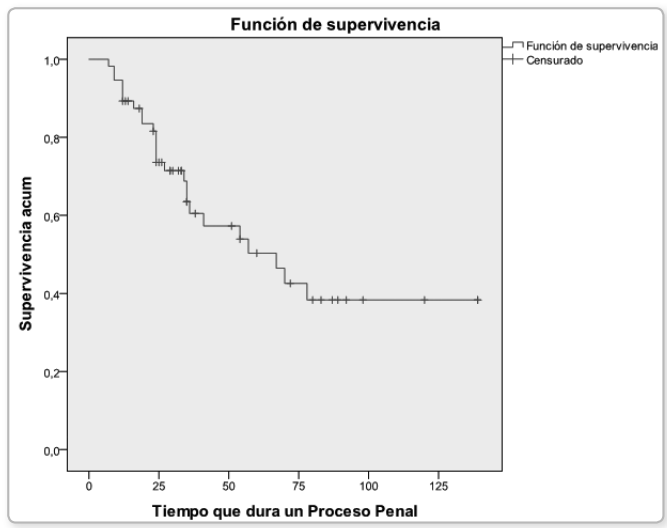
| Media[a] | | | | Mediana | | | |
|---|---|---|---|---|---|---|---|
| | | Intervalo de confianza al 95% | | | | Intervalo de confianza al 95% | |
| Estimaciòn | Error típico | Límite inferior | Límite superior | Estimación | Error típico | Límite inferior | Límite superior |
| 75,987 | 8,510 | 59,307 | 92,667 | 67,000 | 16,515 | 34,632 | 99,368 |

a. La estimaciòn se limita al mayor tiempo de superviviencia, si se ha censurado.

**Percentiles**

| 50,0% | | 75,0% | |
|---|---|---|---|
| Estimación | Error típico | Estimación | Error típico |
| 67,000 | 16,515 | 24,000 | 5,240 |

Luego se muestra la curva de supervivencia. Sitúese en la ordenada la supervivencia acumulada en términos de probabilidad (entre 0 y 1), y el tiempo de supervivencia en el eje de abscisas, permitiéndole responder a preguntas como: ¿Cuántos expedientes sobreviven más de dos años? Trace una línea vertical desde el tiempo 24 meses y en el punto donde corte la curva, lea en horizontal el valor que le corresponde en el eje de ordenadas (daría aproximadamente el 74%).

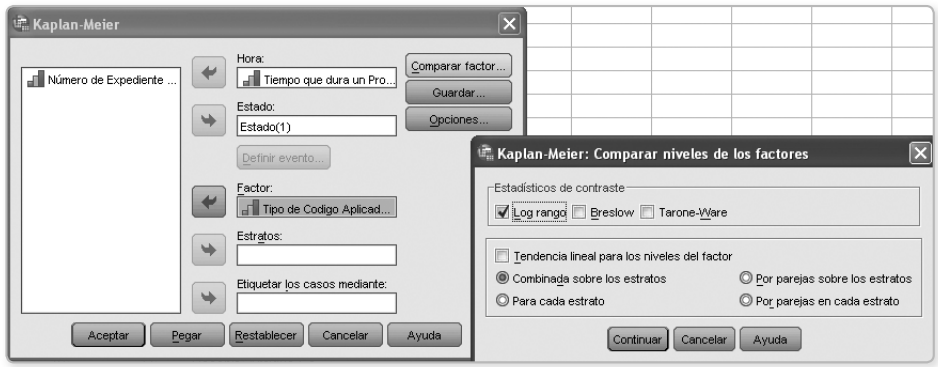



**C. Compare dos o más curvas de supervivencia**

H0: "No hay diferencias" en la supervivencia considerada.

H1: "Si hay diferencias" en la supervivencia considerada.

Analizar > Supervivencia > Kaplan-Meier... y se colocan las dos variables fundamentales en sus respectivas ventanas de selección (Tiempo y Estado). Seguidamente, seleccione la variable que separará dos grupos de proceso y coloque en la ventana Factor; en este ejemplo, la variable *Tipo de código*.




Los resultados con SPSS son los siguientes:

**Resumen del procesamiento de los casos**

| Tipo de Codigo Aplicado al Proceso | Nº total | Nº de eventos | Censurado | |
|---|---|---|---|---|
| | | | Nº | Porcentaje |
| Nuevo Código | 21 | 14 | 7 | 33,3% |
| Antiguo Código | 35 | 11 | 24 | 68,6% |
| Global | 56 | 25 | 31 | 55,4% |

**Tabla de supervivencia**

| Tipo de Codigo Aplicado al Proceso | Tiempo | Estado | Proporción acumulada que sobrevive hasta el momento | | Nº de eventos acumulados | Nº de casos que permanec en |
|---|---|---|---|---|---|---|
| | | | Estimación | Error típico | | |
| Nuevo Código | 7,000 | Concluido | ,952 | ,046 | 1 | 20 |
| | 9,000 | Concluido | ,905 | ,064 | 2 | 19 |
| | 12,000 | Concluido | . | . | 3 | 18 |
| | 12,000 | Concluido | ,810 | ,086 | 4 | 17 |
| | 19,000 | Concluido | ,762 | ,093 | 5 | 16 |
| | 23,000 | Concluido | ,714 | ,099 | 6 | 15 |
| | 24,000 | Concluido | . | . | 7 | 14 |
| | 24,000 | Concluido | . | . | 8 | 13 |
| | 24,000 | Concluido | . | . | 9 | 12 |
| | 24,000 | Concluido | ,524 | ,109 | 10 | 11 |
| | 29,000 | Reserva | . | . | 10 | 10 |
| | 34,000 | Concluido | ,471 | ,110 | 11 | 9 |
| | 41,000 | Concluido | ,419 | ,110 | 12 | 8 |
| | 54,000 | Concluido | ,367 | ,108 | 13 | 7 |
| | 72,000 | Reserva | . | . | 13 | 6 |
| | 78,000 | Concluido | ,306 | ,106 | 14 | 5 |
| | 80,000 | Reserva | . | . | 14 | 4 |
| | 83,000 | Reserva | . | . | 14 | 3 |
| | 92,000 | Reserva | . | . | 14 | 2 |
| | 139,000 | Reserva | . | . | 14 | 1 |
| | 139,000 | Reserva | . | . | 14 | 0 |
| Antiguo Código | 9,000 | Concluido | ,971 | ,028 | 1 | 34 |
| | 12,000 | Concluido | ,943 | ,039 | 2 | 33 |
| | 12,000 | Reserva | . | . | 2 | 32 |
| | 13,000 | Reserva | . | . | 2 | 31 |
| | 14,000 | Reserva | . | . | 2 | 30 |
| | 16,000 | Concluido | ,911 | ,049 | 3 | 29 |
| | 18,000 | Reserva | . | . | 3 | 28 |
| | 19,000 | Concluido | ,879 | ,057 | 4 | 27 |
| | 23,000 | Reserva | . | . | 4 | 26 |
| | 24,000 | Reserva | . | . | 4 | 25 |
| | 25,000 | Reserva | . | . | 4 | 24 |
| | 26,000 | Reserva | . | . | 4 | 23 |
| | 27,000 | Concluido | ,841 | ,066 | 5 | 22 |
| | 29,000 | Reserva | . | . | 5 | 21 |
| | 30,000 | Reserva | . | . | 5 | 20 |
| | 32,000 | Reserva | . | . | 5 | 19 |
| | 33,000 | Reserva | . | . | 5 | 18 |
| | 33,000 | Reserva | . | . | 5 | 17 |
| | 35,000 | Concluido | . | . | 6 | 16 |
| | 35,000 | Concluido | ,742 | ,088 | 7 | 15 |
| | 35,000 | Reserva | . | . | 7 | 14 |
| | 35,000 | Reserva | . | . | 7 | 13 |
| | 35,000 | Reserva | . | . | 7 | 12 |
| | 36,000 | Concluido | ,680 | ,100 | 8 | 11 |
| | 38,000 | Reserva | . | . | 8 | 10 |
| | 51,000 | Reserva | . | . | 8 | 9 |
| | 54,000 | Reserva | . | . | 8 | 8 |
| | 57,000 | Concluido | ,595 | ,118 | 9 | 7 |
| | 60,000 | Reserva | . | . | 9 | 6 |
| | 67,000 | Concluido | ,496 | ,134 | 10 | 5 |
| | 70,000 | Concluido | ,397 | ,139 | 11 | 4 |
| | 87,000 | Reserva | . | . | 11 | 3 |
| | 89,000 | Reserva | . | . | 11 | 2 |
| | 98,000 | Reserva | . | . | 11 | 1 |
| | 120,000 | Reserva | . | . | 11 | 0 |

Medias y medianas del tiempo de supervivencia

| Tipo de Código aplicado al proceso | Media[a] | | | | Mediana | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimaciòn | Error tipico | Intervalo de confianza al 95% | | Estimación | Error típico | Intervalo de confianza al 95% | |
| | | | Límite inferior | Límite superior | | | Límite inferior | Límite superior |
| Nuevo Còdigo | 62,472 | 11,916 | 39,117 | 85,828 | 34,000 | 6,704 | 20,861 | 47,139 |
| Antiguo Código | 74,466 | 9,755 | 55,346 | 93,586 | 67,000 | 8,768 | 49,814 | 84,186 |
| | 75,987 | 8,510 | 59,307 | 92,667 | 67,000 | 16,515 | 34,632 | 99,368 |

a. La estimaciòn se limita al mayor tiempo de superviviencia si se ha censurado.

**Percentiles**

| Tipo de Codigo Aplicado al Proceso | 50,0% | | 75,0% | |
|---|---|---|---|---|
| | Estimación | Error típico | Estimación | Error típico |
| Nuevo Código | 34,000 | 6,704 | 23,000 | 4,140 |
| Antiguo Código | 67,000 | 8,768 | 35,000 | 4,919 |
| Global | 67,000 | 16,515 | 24,000 | 5,240 |

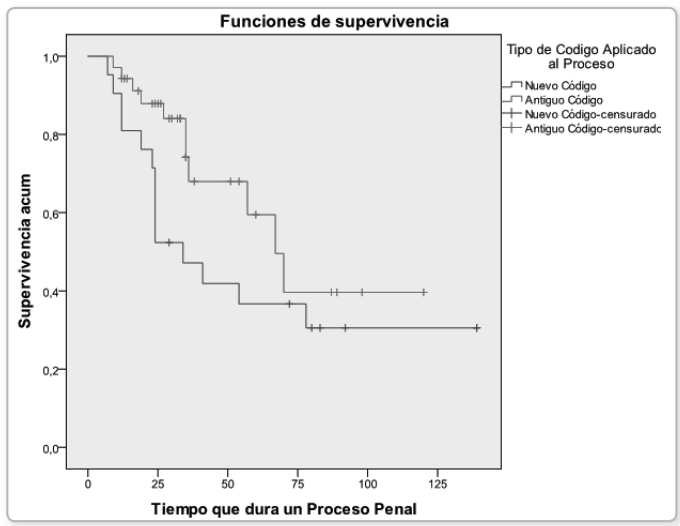
Log Rank, con el valor del estadístico (2,68), los grados de libertad (1, en este caso, al ser dos grupos los que se comparan) y el valor "p" asociado al contraste (0,1018) que corresponde a una probabilidad que no le permite rechazar la hipótesis nula; es decir, no hay diferencias estadísticamente significativas en la supervivencia de ambos grupos.

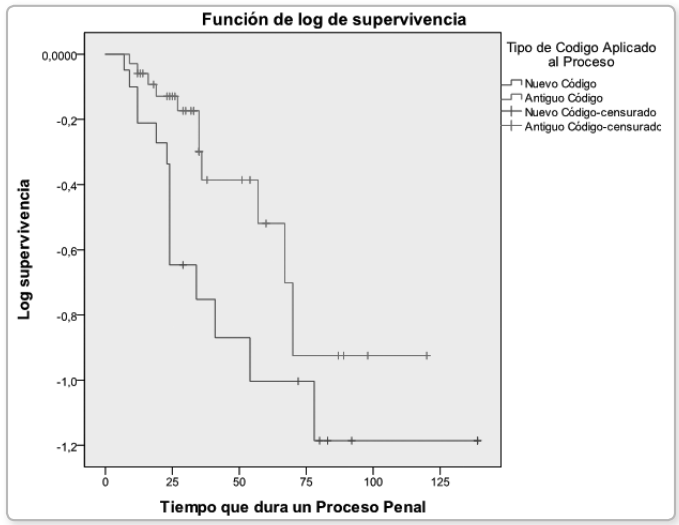
**Comparaciones globales**

| | Chi-cuadrado | gl | Sig. |
|---|---|---|---|
| Log Rank (Mantel-Cox) | 2,678 | 1 | ,102 |

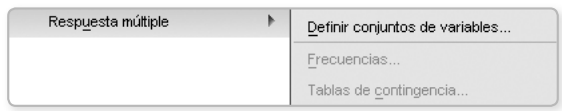
Prueba de igualdad de distribuciones de supervivencia para diferentes niveles de Tipo de Código Aplicado al Proceso.

Como puede apreciarse, hay una separación entre ambas curvas de supervivencia a favor del grupo del código antiguo (parecen sobrevivir más), pero el test estadístico Log Rank ya se ha visto que no alcanza el valor crítico para rechazar la $H_0$ de "no hay diferencias" en la supervivencia globalmente considerada. Si se observa con mayor detalle la figura, se comprueba como al principio (entre el primero y el quinto año), que sí parecen separarse las curvas de supervivencia, aunque no hay demasiada diferencia, a partir de los 60 meses de seguimiento.

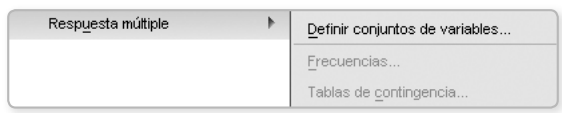





## 9.3. RESPUESTAS MÚLTIPLES




SPSS permite obtener frecuencias y porcentajes para variables que son múltiples. Una variable múltiple es aquella variable que puede tomar más de un valor para cada caso.




### A. Definir conjuntos

Agrupa variables elementales en conjuntos de categorías múltiples y de dicotomías múltiples, para los que se pueden obtener tablas de frecuencias y tablas de contingencia. Se pueden definir hasta 20 conjuntos de respuestas múltiples. Cada conjunto debe tener un nombre exclusivo. Las variables elementales se pueden codificar como dicotomías o categorías.

- **Dicotomías.** En Valor contado debe introducirse un entero. Cada variable que tenga al menos una aparición del valor contado, se convierte en una categoría del conjunto de dicotomías múltiples.
- **Categorías.** Crear un conjunto de categorías múltiples con el mismo rango de valores que las variables que lo componen. Deben introducirse dos enteros como valores máximo y mínimo del rango para las categorías del conjunto de categorías múltiples. El procedimiento suma cada valor entero distinto en el rango inclusivo para todas las variables que lo componen. Las categorías vacías no se tabulan, y a cada conjunto de respuestas múltiples se le debe asignar un nombre exclusivo de hasta siete caracteres. El procedimiento coloca delante del nombre asignado un signo dólar ($). El nombre del conjunto de respuestas múltiples solo se encuentra disponible para su uso en los procedimientos de respuestas múltiples.

**B. Frecuencias**

Produce tablas de frecuencias para conjuntos de respuestas múltiples. Es necesario definir previamente uno o más conjuntos de respuestas múltiples. Para los conjuntos de dicotomías múltiples, los nombres de categorías que se muestran en los resultados, proceden de etiquetas de variables definidas para variables elementales del grupo.

Para los conjuntos de categorías múltiples, las etiquetas de categoría proceden de las etiquetas de valor de la primera variable del grupo. Si las categorías perdidas para la primera variable están presentes para otras variables del grupo, debe definirse una etiqueta de valor para las categorías perdidas.

Los casos con Valores perdidos se excluyen sobre la base de la tabla por tabla. Un caso se considera perdido para un conjunto de dicotomías (o categorías múltiples) solo si alguno de sus componentes contiene el valor (o tiene valores válidos en el rango). Se pueden excluir los casos con valores perdidos en cualquier variable de la tabulación del conjunto de dicotomías o categorías, ambas.

**C. Tablas de contingencia**

Presenta, en forma de tabla de contingencia, conjuntos de respuestas múltiples, variables elementales o una combinación. También permite obtener porcentajes de casilla basados en casos o respuestas, modificar la gestión de los valores perdidos u obtener tablas de contingencia emparejadas.

Es necesario definir previamente uno o más conjuntos de respuestas múltiples. Los nombres en los resultados se muestran de la misma forma que en el procedimiento anterior, y las etiquetas de categoría por columnas se muestran en tres líneas, con un máximo de ocho caracteres por línea. Para evitar la división de palabras, se pueden invertir los elementos de las filas y las columnas o volver a definir las etiquetas.

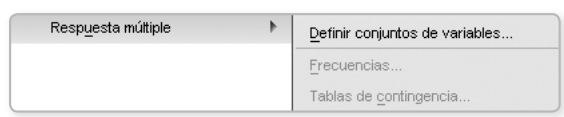
### 9.3.1. EJERCICIO PRÁCTICO 1

El ejercicio permitirá utilizar los principales procedimientos del menú Clasificar. Los datos usados son del siguiente ejemplo.

Suponga que existen cinco terapias [Terapia 1 (1); Terapia 2 (2); Terapia 3 (3); Terapia 4 (4) y Terapia 5 (5)] para pacientes con cáncer que se le suministra uno por día. Cada paciente debe elegir tres terapias como máximo. Además, debe seleccionar los días de la semana [lunes (1); martes (2); miércoles (3); jueves (4) y viernes (5)] en que recibirá la terapia. Los datos se encuentran en el archivo *RespuestasMiltiples.sav.*

Dentro del menú **Analizar** se encuentra el procedimiento **Respuestas múltiples**. El aspecto de dicho menú es el siguiente:

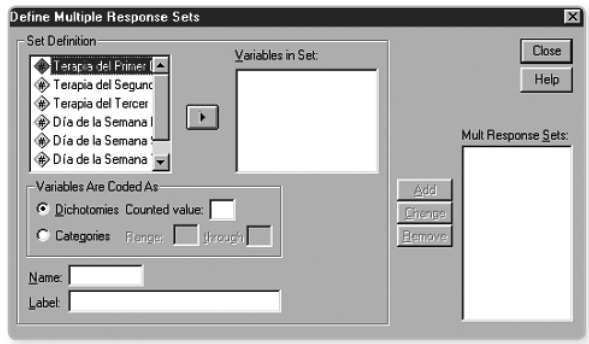
Analizar > Respuestas múltiples (Múltiple Response)




Para realizar el procedimiento Respuestas múltiples, se deben seguir los cuadros de diálogo que se presentan. A continuación, se muestra una explicación detallada del cuadro de diálogo correspondiente a los principales procedimientos.

### A. Definir conjuntos

Para definir conjuntos, seleccione la opción Analizar > Respuestas múltiple > Definir conjuntos, obteniéndose el cuadro de diálogo siguiente:



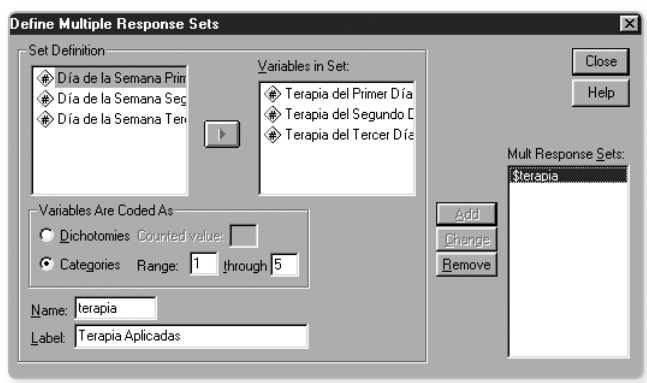

Primero, defina cuáles son las variables que van a formar parte de la variable múltiple. El cuadro de la izquierda muestra la lista de variables del archivo. Las variables seleccionadas se trasladan al cuadro del lado derecho «Campo variables de conjuntos» (Variables in Set:). En este caso, las variables son Terapia del primer día, Terapia del segundo día y Terapia del tercer día (ver fig.).

En el campo **Las variables están codificados como** (Variables Are Coded As), indique la forma cómo se han introducido las variables que van a formar parte de la variable Multirrespuesta; es decir, si se han codificado como dicotómicas o con tantas categorías como posibles respuesta tenga. En el ejemplo, son variables con cinco categóricas cada una.

En el campo **Nombre** (Name), se especifica el nombre que se va dar a la variable múltiple; en el ejemplo es Terapia.

En el campo Etiqueta (Label), se especifica una etiqueta para dicha variable: Terapias aplicadas.

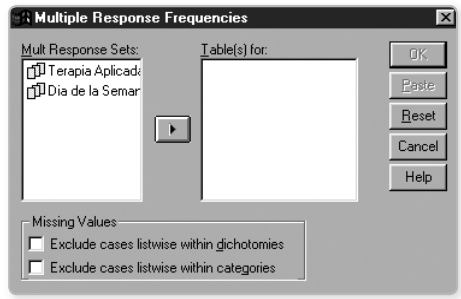
Una vez que se han rellenados estos campos presione el botón Añadir (Add), para que la variable que se acaba de definir figure en el campo de Conjunto de respuestas múltiples (Mult Response Sets).



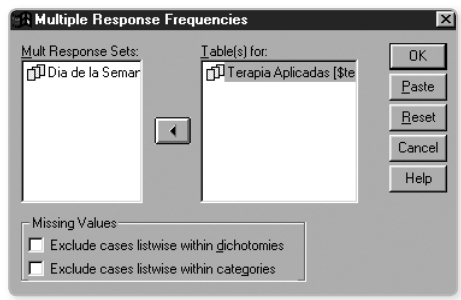

Una vez que se han definido las variables múltiples presione Cerrar (Close). Entonces, ya se pueden obtener las frecuencias de dichas variables o cruzar variables de este tipo.

### B. Frecuencias

Para obtener frecuencias y porcentajes de una o más variables múltiples, se selecciona la opción Analizar > Respuestas múltiple > Frecuencias, se obtiene el cuadro de diálogo siguiente:




En el campo **Conjunto de respuestas múltiples** (Mult Response Sets), figuran los nombres de las variables múltiples que previamente se han definido. De esta lista, se selecciona las que se quieren analizar y se traslada a las **Tablas para** (Table[s] For). En este ejemplo, halle frecuencias Terapia aplicada:




Para este tipo de análisis, SPSS considera como missing aquellos casos que tienen valor missing en todas las variables que forman la variable múltiple. Si en el campo Valores perdidos se selecciona alguna de las dos opciones: *Excluir casos según lista en las dicotómicas o Excluir casos según lista en las categorías*. SPSS excluye del análisis aquellos casos que tienen valor missing en alguna de las variables que forman la variable múltiple.

Si se realiza el análisis de frecuencias de la variable múltiple Terapias aplicadas, que está formada por tres variables (Terapia del primer día, Terapia del segundo día y Terapia del tercer día). El resultado se muestra en el siguiente cuadro:

```
—
Group $TERAPIA  Terapia Aplicadas

                                          Pct of  Pct of
Category label           Code      Count  Responses  Cases

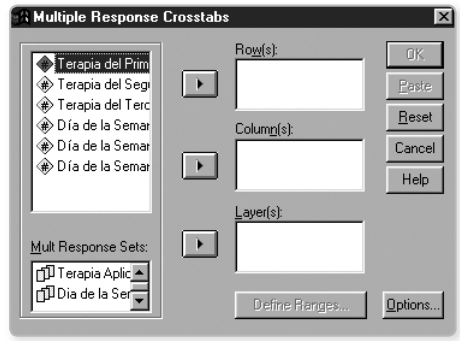
Terapia 1                   1         20     22.2     66.7
Terapia 2                   2         19     21.1     63.3
Terapia 3                   3         17     18.9     56.7
Terapia 4                   4         21     23.3     70.0
Terapia 5                   5         13     14.4     43.3
                                  -------    -----    -----
               Total responses        90    100.0    300.0

0 missing cases;  30 valid cases
```

Esta tabla contiene las frecuencias absolutas y los porcentajes calculados sobre el total de las respuestas (90).

### C. Tablas de contingencia

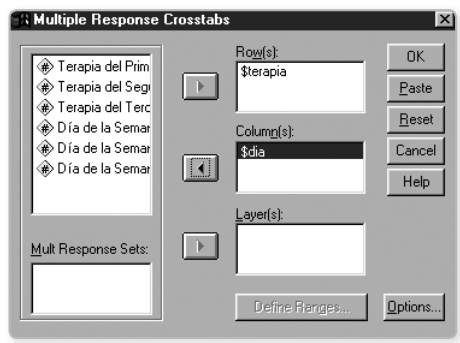
Para obtener tablas de contingencia de una o más variables múltiples, se selecciona la opción **Analizar > Respuestas múltiple > Tablas de contingencia** (Crosstabs), obteniéndose el cuadro de diálogo siguiente:



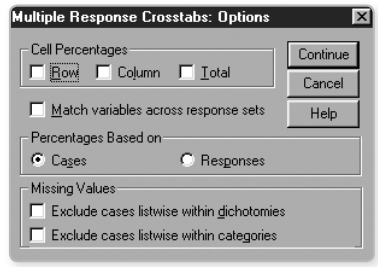

En el campo **Filas** (Row[s]), se incluyen las variables que se van a representar en las filas de las tablas de frecuencia, y en el campo **Columnas** (Column[s]), las variables que se van a representar las columnas de dichas tablas. En ambos campos se pueden incluir variables múltiples; por tanto, se va a obtener una tabla de frecuencias observadas, cruzando la primera variable detallada en el campo **Filas** con la primera variable especificada en el campo **Columna**, luego se obtiene otra tabla cruzando la primera variable especificada en el campo **Filas** con la segunda variable especificada en el campo **Columna,** y así sucesivamente con todas las variables especificadas en el campo **Columnas**. Después se realizarán las tablas cruzando la segunda variable del campo **Filas** con cada una de las del campo **Columna,** y así sucesivamente.

Las tablas anteriores se pueden obtener en función de los valores que tomen otras variables que se especificarán en el campo **Capas** (Layer[s]).

En el ejemplo, se hallarán las tablas de contingencia para la variable múltiple Terapia aplicada y Día de la semana:




Si pulsa el botón **Opciones** (Options...), se obtiene una ventana que contiene la lista de resultados que se pueden obtener además de las frecuencias observadas; es decir, el cuadro de diálogo siguiente:




Si se realiza el análisis de tablas de contingencia de la variable múltiple Terapias aplicadas, que está conformada por tres variables (Terapia del primer día, Terapia del segundo día y Terapia del tercer día) y la variable Día de la semana que está formada por tres variables (Día 1, Día 2 y Día 3). El resultado se muestra en el siguiente cuadro:

* * * C R O S S T A B U L A T I O N * * *

$TERAPIA (group) Terapia Aplicadas
by $DIA (group) Dia de la Semana

| | | $DIA | | | | | |
|---|---|---|---|---|---|---|---|
| | Count | Lunes | Martes | Miercole | Jueves | Viernes | Row Total |
| | | 1 | 2 | 3 | 4 | 5 | |
| $TERAPIA | | | | | | | |
| Terapia 1 | 1 | 14 | 15 | 11 | 13 | 7 | 20<br>66.7 |
| Terapia 2 | 2 | 13 | 12 | 11 | 14 | 7 | 19<br>63.3 |
| Terapia 3 | 3 | 13 | 11 | 9 | 13 | 5 | 17<br>56.7 |
| Terapia 4 | 4 | 15 | 13 | 9 | 18 | 8 | 21<br>70.0 |
| Terapia 5 | 5 | 8 | 9 | 5 | 11 | 6 | 13<br>43.3 |
| | Column Total | 21<br>70.0 | 20<br>66.7 | 15<br>50.0 | 23<br>76.7 | 11<br>36.7 | 30<br>100.0 |

Percents and totals based on respondents

30 valid cases; 0 missing cases

En esta tabla, se muestra la distribución conjunta de las dos variables: múltiple Terapia aplicada y la variable Día de la Semana. Las filas corresponden a Terapia aplicada y las columnas son Días de la semana. Los valores que aparecen en cada celda corresponden a frecuencias observadas. Además, se pueden calcular las Porcentajes por filas, Porcentajes por columnas, Porcentajes totales.

CAPÍTULO

10

# GRÁFICOS

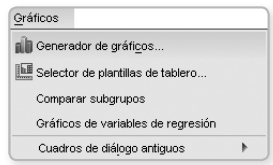
El menú Gráficos permite realizar una gran cantidad de gráficos orientados al análisis de los datos. La primera opción Galería, no realiza un gráfico en sí, sino que se traslada directamente a un tema de la ayuda que presenta una galería de gráficos. La segunda opción es la de gráficos Interactivos. Con esta opción se generan gráficos que podrán ser modificados posteriormente, en cuanto a las variables y elementos gráficos, de forma mucho más efectiva que los gráficos normales. Sin embargo, solo son convenientes cuando se trabaje con archivos de datos pequeños o subconjuntos de otros archivos mayores, ya que los requerimientos de recursos son enormes. Las restantes opciones de este menú son una lista de todos los gráficos posibles (no interactivos).

## 10.1. GALERÍA DE GRÁFICOS

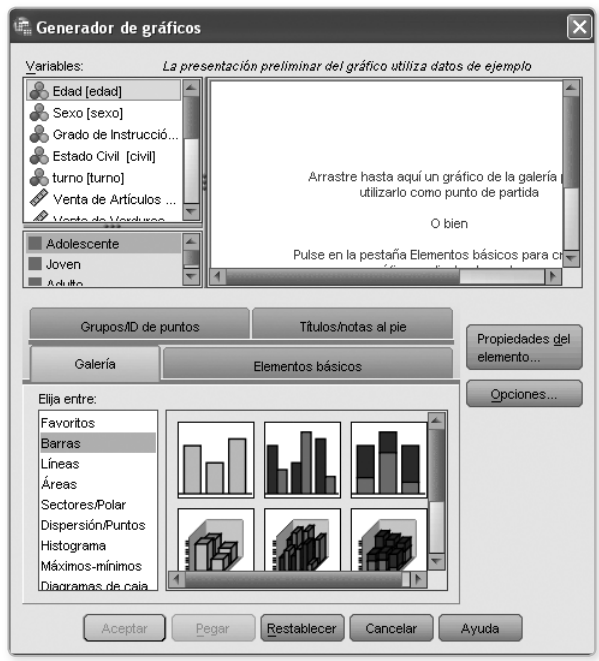
El menú Gráficos permite realizar una gran cantidad de gráficos orientados al análisis de los datos (Análisis > Gráficos).




Si ingresa a Generador de gráficos.




Muestra el siguiente cuadro de diálogo:



### A. Barras

Para la creación de un gráfico de barras, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, se pasa al cuadro de diálogo (ver fig.). Las posibilidades en los gráficos de barras son las siguientes:

a. **Simples (**  **)**. Las barras representan una función de la variable o variables.

b. **Agrupadas (**  **)**. Las barras representan una función de la variable o variables, dentro de categorías de otra variable.

c. **Apiladas (**  **)**. Las barras representan una función de la variable o variables, dentro de categorías de otra variable y se apilan por segmentos. Independientemente de cada posibilidad, se tiene que:
   - **Resúmenes para grupos de casos:** Resume categorías de una sola variable.
   - **Resúmenes para distintas variables:** Resume categorías de dos o más variables.
   - **Valores individuales de los casos:** Resume una sola variable por cada caso, con lo cual se tienen nueve posibilidades.
     - **Resúmenes para grupos de casos en gráficos de barras simples:** Cree un gráfico que resuma las categorías de una sola variable. La altura de la barra representa una función de la propia variable de categoría u otra variable de resumen.
     - **Resúmenes para distintas variables en gráficos de barras simples:** Crea un gráfico que resume dos o más variables. Hay una barra por cada variable.
   - **Valores individuales de los casos en gráficos de barras simples:** Crea un gráfico que resume una sola variable por cada caso.
     - **Resúmenes para grupos de casos en gráficos de barras agrupadas:** Crea un gráfico que resume categorías de una variable dentro de categorías de otra variable. La altura de la barra puede representar una función de la variable para la definición de las agrupaciones o bien otra variable de resumen.
     - **Resúmenes para distintas variables en gráficos de barras agrupadas:** Crea un gráfico que resume dos o más variables dentro de categorías de otra variable.
   - **Valores individuales de los casos en gráficos de barras agrupadas.** Crea un gráfico que resume valores de dos o más variables por cada caso.
     - **Resúmenes para grupos de casos en gráficos de barras apiladas:** Crea un gráfico en el cual los segmentos de las barras se apilan uno encima de otro. Hay una barra por cada valor de la variable del eje de categorías. Los segmentos de cada barra representan las categorías de la variable de pila.
     - **Resúmenes para distintas variables en gráficos de barras apiladas:** Crea un gráfico que resume dos o más variables dentro de categorías.
   - **Valores individuales de los casos en gráficos de barras apiladas:** Crea un gráfico que resume dos o más variables para cada caso.

### B. Líneas

Para la creación de un gráfico de barras, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido dicho gráfico, revise el cuadro de diálogo usual que aparece cuando se realiza un análisis. Las posibilidades en los gráficos de líneas son las siguientes:

a. **Simples (**  **)**. Igual que los de barras.

b. **Múltiples (**  **)**. Igual que los de barras agrupadas.

c. **Líneas verticales (**  **)**. Igual que el anterior.

Independientemente de cada posibilidad, se tiene que:

- Resúmenes para grupos de casos
- Resúmenes para distintas variables
- Valores individuales de los casos

Que presenta las mismas opciones que los gráficos de barras ya comentados.

### C. Áreas

Para la creación de un gráfico de áreas, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, revise el cuadro de diálogo usual que aparece cuando se realiza un análisis.

Las posibilidades en los gráficos de áreas son las siguientes:

a. **Simples (**  **)**. Igual que los de barras.

b. **Apiladas (**  **)**. Igual que los de barras agrupadas.

Independientemente de cada posibilidad, se tiene que:

- Resúmenes para grupos de casos
- Resúmenes para distintas variables
- Valores individuales de los casos

### D. Sectores

Para la creación de un gráfico de sectores, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido dicho gráfico revise el cuadro de diálogo usual que aparece cuando se realiza un análisis.

 La buena percepción visual de los gráficos de sectores se basa en su sencillez, que por otra parte, imposibilita mostrar variables de acuerdo con categorías de otras variables, con lo cual no hay distintos gráficos posibles, solo se puede determinar la estructura de los datos como se ha visto:

- Resúmenes para grupos de casos
- Resúmenes para distintas variables
- Valores individuales de los casos

### E. Máximos y mínimos

Para la creación de un gráfico de máximos y mínimos, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión revise el cuadro de diálogo usual que aparece cuando se hace un análisis.

Las posibilidades en los gráficos de máximos y mínimos son las siguientes:

a. **Máximo, mínimo y cierre simple (  ).** Crea un gráfico que resume una variable dicotómica (de dos valores distintos) o tricotómica (de tres valores distintos) en categorías de otra variable.

b. **Máximo, mínimo y cierre agrupado (  ).** Crea un gráfico que resume dos o tres variables en agrupaciones definidas por una variable categórica. Cada barra de una agrupación se define mediante una segunda variable categórica.

c. **Línea de diferencias (  ).** Crea un gráfico que resume una o dos variables dicotómicas en categorías de otra variable.

d. **Barras de rango simples (  ).** Crea un gráfico que resume una o dos variables numéricas en categorías de otra variable.

e. **Barras de rango agrupadas (  ).** Crea un gráfico que resume dos o más parejas de variables, en agrupaciones definidas por una variable categórica. Cada barra de una agrupación representa una de las parejas de variables. Para los máximos y mínimos, se tienen las opciones ya vistas:
   - Resúmenes para grupos de casos
   - Resúmenes para distintas variables
   - Valores individuales de los casos

### F. Pareto

Para la creación de un gráfico de Pareto, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, revise el cuadro de diálogo usual que aparece al realizar un análisis. Se tienen las posibilidades gráficas:

- **Simple (  ).** Crea un gráfico de barras que resume categorías de una variable o varias (según la opción seleccionada), ordenadas en sentido descendente. Una línea muestra la suma acumulada.
- **Apilado (  ).** Crea un gráfico de barras que resume categorías de una variable o varias en categorías de otra, con las barras ordenadas en sentido descendente. Una línea representa la suma acumulada. De la misma forma que en los gráficos anteriores, se tiene que:
  - Recuentos o sumas para grupos de casos
  - Sumas para distintas variables
  - Valores individuales de los casos

### G. Control

Para la creación de un gráfico de control, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión se pasa al cuadro de diálogo usual que aparece cuando se hace un análisis. Los gráficos de control de procesos ofrecen los subtipos:

- **X-barra, R, s (  ).** Crea gráficos X-barra, R y s que representan la media, el rango y la desviación típica, respectivamente, en medidas de algunos procesos en ejecución (a lo largo del tiempo).
- **Individuos, rango móvil (  ).** Representa una única variable de medida del proceso por caso individual. En los gráficos de rango móvil, cada punto representa la diferencia entre el valor actual y el anterior.

- **p, np (  ).** Resume la proporción o el número de unidades disconformes en los subgrupos.
- **c, u (  ).** Crea un gráfico de control de proceso que resume el número de disconformidades (o disconformidades por unidad) dentro de los subgrupos. Las opciones posibles son las siguientes:
  - **Los casos son unidades.** Se resume una variable de proceso.
  - **Los casos son subgrupos.** Se resumen las variables en subgrupos de casos definidos por una variable categórica. Cada punto resume las variables seleccionadas en un único caso. Esta opción no está disponible para los gráficos de control de *Individuos, rango móvil*.

### H. Diagramas de cajas

Para la creación de un gráfico de diagrama de cajas, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión se pasa al cuadro de diálogo usual que aparece al hacer un análisis. Existen dos tipos de diagramas:

- **Simple (  ).** Crea un diagrama de caja que resume una variable numérica en categorías de otra variable o bien resume varias variables. Cada caja muestra la mediana, los cuartiles y los valores extremos, según un grupo de casos o según una variable.
- **Agrupado (  ).** En este caso, se representa dentro de las agrupaciones definidas por una variable categórica. Cada caja de una agrupación se define mediante una segunda variable categórica y dos opciones ya vistas:
  - Resúmenes para grupos de casos
  - Resúmenes para distintas variables

### I. Barras de error

Para la creación de un gráfico de barras de error, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido dicho gráfico, se pasa al cuadro de diálogo usual que aparece cuando se realiza un análisis. Presenta dos posibilidades como en el caso anterior:

- **Simples (  ).** Crea un gráfico que resume la distribución de una única variable numérica en categorías de otra variable, o resume varias variables numéricas.
- **Agrupadas (  ).** Crea un gráfico que resume la distribución de una o más variables numéricas en agrupaciones definidas por una variable categórica. Igual que en el caso anterior:
  - Resúmenes para grupos de casos
  - Resúmenes para distintas variables

### J. Dispersión

Para la creación de un gráfico de dispersión, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido dicho gráfico se revisa el cuadro de diálogo usual que aparece al momento de hacer el análisis. Este tipo de gráficos son más específicos en contraste con los anteriores. Solo se tiene cuatro opciones sin más posibilidades:

- **Simple (  ).** Crea un gráfico de dispersión en dos dimensiones que relaciona entre sí dos variables numéricas. Cada variable se representa en un eje.

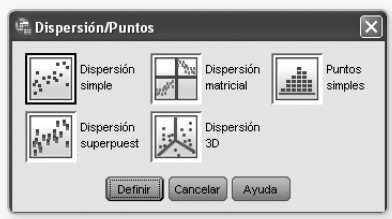
## Ejemplo 1

Haga uso de la herramienta Gráficos para poder explicar con claridad este modelo. La primera aproximación a la relación entre dos variables (*X* e *Y*) puede hacerse a partir de un diagrama de dispersión. En este caso, se tienen los siguientes datos:

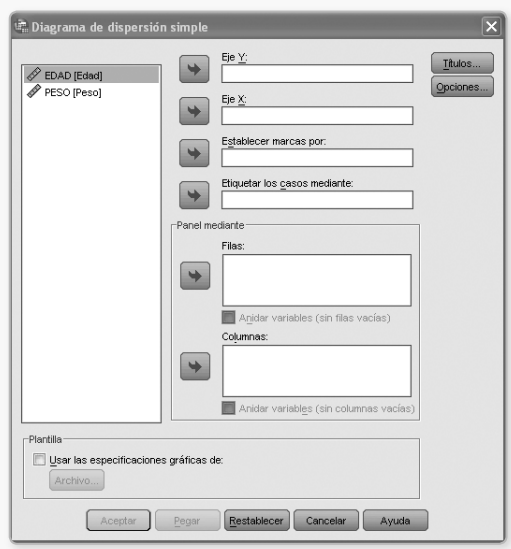
| Sujetos | Edad | Peso |
|---|---|---|
| 1 | 39 | 58 |
| 2 | 30 | 55 |
| 3 | 50 | 65 |
| 4 | 52 | 70 |
| 5 | 62 | 69 |

Los datos corresponden a cinco sujetos medidos en dos variables: Edad y Peso. Como en otros ejemplos previos, introduzca los datos correspondientes a los cinco sujetos en las dos variables, en un archivo nuevo del Editor de datos. Una vez introducidos los datos guárdelos como *regresion.sav*.

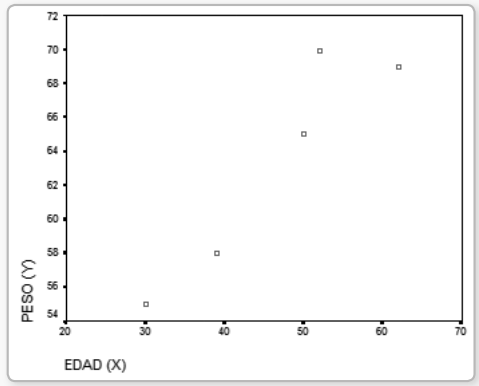
Para realizar el diagrama de dispersión, se selecciona el menú Gráficos de la barra de menús del Editor de datos y se elige el procedimiento Dispersión (Scatter). El cuadro de diálogo correspondiente es el siguiente:




Luego pulse Definir, para definir los ejes del diagrama de las variables, obteniendo el cuadro de diálogo siguiente:



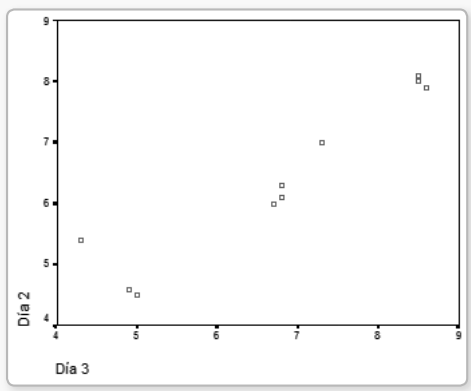
Con esta definición, se obtiene el siguiente diagrama de dispersión simple en el Visor de resultados:




Como se observa, la forma del diagrama indica que los puntos no están perfectamente alineados, pero se acercan a una hipotética línea recta.

## Ejemplo 2

La correlación y regresión lineales permiten evaluar la existencia de relaciones lineales entre variables. Sin necesidad de calcular algún estadístico, es posible elaborar un gráfico de dispersión que informe sobre la posibilidad de este tipo de relación. En este caso, mediante el procedimiento **Gráficos** > **Dispersión** > **Simples** es posible elaborar el siguiente gráfico de dispersión de las variables *Día 2* y *Día 3*:



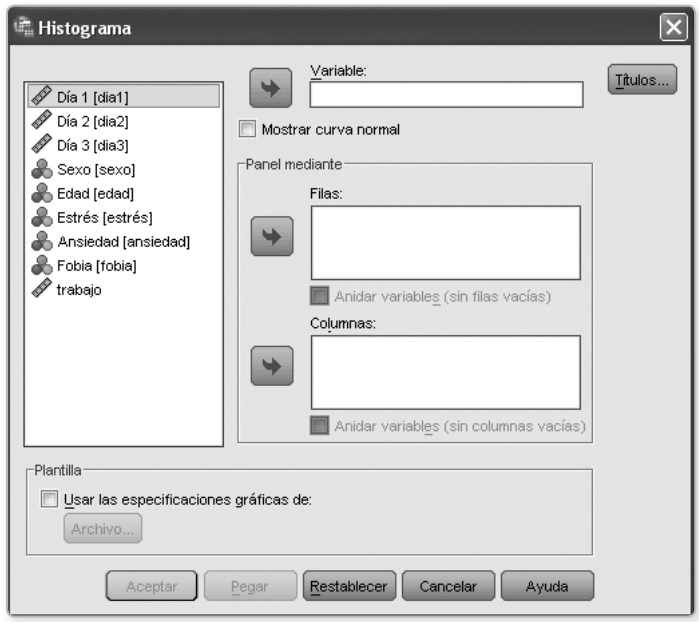

Cada punto del gráfico se corresponde con las puntuaciones de un sujeto en ambas variables. El gráfico sugiere que existe una relación lineal directa (o positiva) entre ambas, de modo que los sujetos que trabajan más el segunda día también lo hacen el tercer día.

Para comprobar si esta relación es estadísticamente significativa, se utilizan los procedimientos **Correlaciones** y **Regresión lineal**.

- **Superpuesto ( ).** Crea un gráfico con dos o más pares de variables. Representa simultáneamente los valores de la primera variable de cada par frente a los de la segunda.
- **Matricial ( ).** Crea un diagrama con todas las combinaciones posibles de dos o más variables numéricas entre sí. Es una representación en dos dimensiones, de los valores de cada una de las variables de matrices frente a los de cada una de las restantes.
- **3D ( ).** Crea un gráfico de tres variables numéricas en tres dimensiones. Representa los valores de una variable frente a los de las otras dos.

### K. Histograma

Para la creación de un gráfico de histograma, se muestra el cuadro de diálogo inicial de selección del gráfico deseado.




Una vez elegido el gráfico en cuestión, se pasa al cuadro de diálogo usual que aparece cuando se realiza un análisis.

Crea un histograma que muestra la distribución de una única variable numérica. Un histograma representa la distribución de una variable cuantitativa que muestra la concentración relativa de los datos, a lo largo de diferentes intervalos o secciones de la escala en la que están medidos dichos datos.

### L. P-P

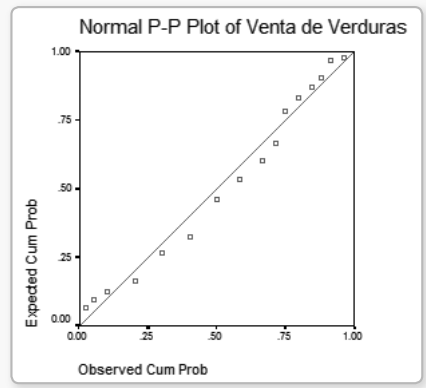
Para la creación de un gráfico de P-P, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido dicho gráfico se revisa el cuadro de diálogo usual que aparece cuando se hace un análisis.

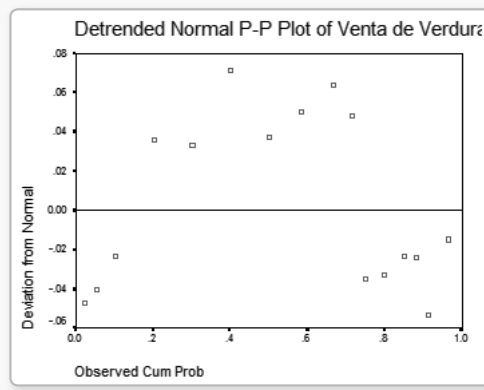
 Crea un gráfico de las proporciones acumuladas de una variable respecto a las de una distribución cualquiera de prueba. Es decir, por cada variable especificada, proporciona dos gráficos: en el primero, se representan los valores de la función de distribución acumulativa esperada bajo el supuesto de normalidad (por defecto) frente a los observados; en el segundo, se representan los residuos.

Los gráficos de probabilidad suelen emplearse para determinar si la distribución de una variable coincide con una distribución dada. Si la variable seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta. Entre las distribuciones de prueba disponibles, se encuentran beta, chi-cuadrado, exponencial, gamma, seminormal, Laplace, logístico, Lognormal, normal, Pareto, t de Student, Weibull y uniforme. Según la distribución elegida, pueden especificarse distintos grados de libertad y otros parámetros.

**Ejemplo:**

Para la variable venta de verduras, se puede decir que sí tiene una distribución normal.



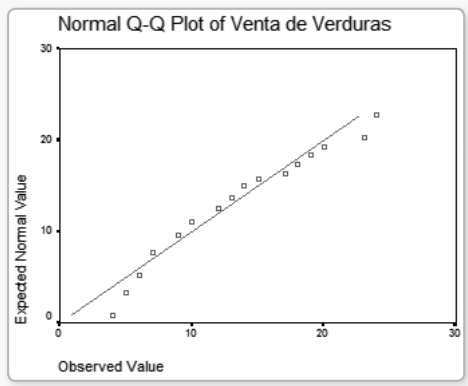
### M. Q-Q

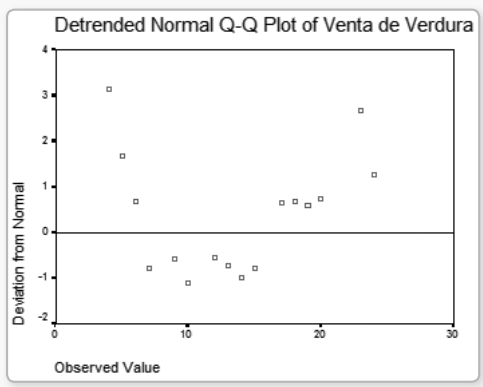
Para la creación de un gráfico de Q-Q, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, se revisa el cuadro de diálogo usual que aparece al momento de hacer un análisis.

 Crea un gráfico con los cuantiles de distribución de una variable respecto a los cuantiles de una distribución cualquiera de prueba; es decir, por cada variable especificada proporciona dos gráficos: en el primero, se representan los valores de los cuantiles esperados bajo el supuesto de distribución normal (por defecto) de media cero y varianza uno frente a los observados; en el segundo, en lugar de los cuantiles esperados, se representan los residuos. Las distribuciones de prueba disponibles son las mismas que en P-P.

**Ejemplo:**

Para la variable Venta de verduras, se puede decir que sí tiene una distribución normal.



**N. Secuencia**

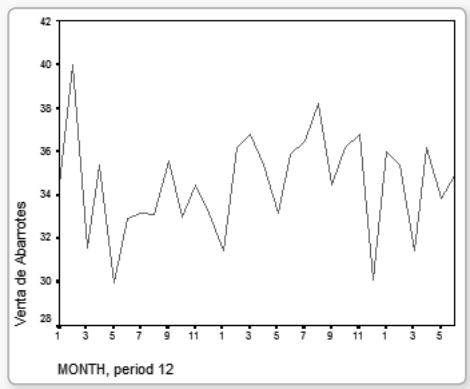
Para la creación de un gráfico de secuencias, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, se pasa al cuadro de diálogo usual que aparece al realizar un análisis.

Crea un gráfico de casos en secuencia. Este procedimiento requiere datos de serie temporal u otros en los que se ordenen los casos con un criterio significativo. Representa, en dos dimensiones, los valores de cada una de las variables especificadas frente al número de secuencia en el archivo de datos.

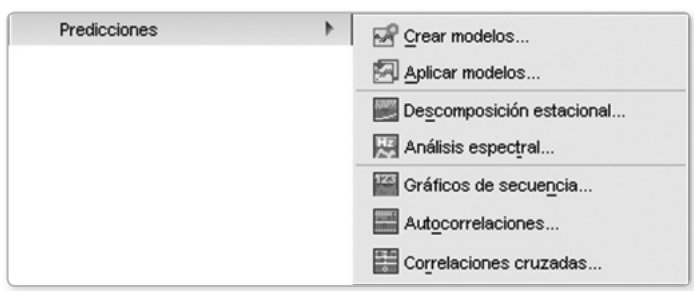
Como se ha dicho anteriormente, si la variable es una serie temporal a cuyos valores se le han asignado fechas (opción Definir fechas), la representación se realiza frente al instante de tiempo correspondiente.

**Ejemplo:**

Para el cuadro de diálogo anterior, el gráfico es el siguiente:




## 10.2. SERIE TEMPORAL



### 10.2.1. SERIE TEMPORAL: AUTOCORRELACIONES

Para la creación de un gráfico de serie temporal, autocorrelaciones, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, se pasa al cuadro de diálogo usual que aparece cuando se realiza un análisis.

 Crea un gráfico con la función de autocorrelación simple (FAS) y la función de autocorrelación parcial de una o varias series. Este procedimiento solo es adecuado en datos de serie temporal.

- **Autocorrelación simple.** Correlaciona los valores de una serie con los valores retardados en uno o más casos. Las autocorrelaciones se calculan para los retardos de orden 1, 2... hasta un número especificado.
- **Autocorrelación parcial.** Correlaciona los valores de una serie con los valores retardados en uno o más casos, después de haber eliminado los efectos de las correlaciones para los retardos intercalados.

### 10.2.2. SERIE TEMPORAL: CORRELACIONES CRUZADAS

Para la creación de un gráfico de serie temporal, correlaciones cruzada, se presenta el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido dicho gráfico se pasa al cuadro de diálogo usual que aparece cuando se hace un análisis.

-  Crea un gráfico con la función de correlación cruzada de dos o más series para retardos positivos, negativos y de valor cero; es decir, representa para cada par de variables especificadas, los valores de la función de correlación cruzada. Este procedimiento solo es apropiado en datos de serie temporal.
- **Correlación cruzada**. Correlación que se da entre dos series temporales. Las observaciones de una serie se correlacionan con las observaciones de otra serie para diversos retardos y adelantos.

Se representan habitualmente en un gráfico que ayuda a identificar las variables que son indicadores precedentes de otras variables. El procedimiento CCF, del módulo *Tendencias*, genera las correlaciones cruzadas.

### 10.2.3. SERIE TEMPORAL: ANÁLISIS ESPECTRAL

Para la creación de un gráfico de serie temporal: análisis, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión se pasa al cuadro de diálogo usual que aparece al realizar un análisis. El análisis espectral analiza la variación de las series como un todo en componentes periódicos de frecuencia diferente, en lugar de analizar la variación desde un punto temporal al siguiente, que es lo usual en las técnicas de serie temporal.

Las series suaves tienen más variación a bajas frecuencias, la variación aleatoria ("ruido blanco") dispersa la variación a través de todas las frecuencias. Este análisis representa para cada variable especificada los valores del periodograma (u otra función de análisis espectral) con objeto de obtener una estimación de densidad espectral. Para suavizar el periodograma, se dispone de Tukey-Hamming, Tukey, Parzen, Bartlett, Unidad de Daniell o de ninguno. Las series con datos perdidos no pueden analizarse con este procedimiento.

## 10.3. COR

Para la creación de un gráfico de COR, se muestra el cuadro de diálogo inicial de selección del gráfico deseado. Una vez elegido el gráfico en cuestión, revise el cuadro de diálogo usual que aparece al hacer un análisis. Este procedimiento es un método útil para evaluar la realización de esquemas de clasificación en los que exista una variable con dos categorías por las que se clasifiquen los sujetos.

Permite calcular la estimación del área situado bajo la curva COR, de forma paramétrica o no paramétrica, mediante un modelo exponencial binegativo. Además, permite especificar si se debe incluir o excluir el valor del punto de corte al realizar una clasificación positiva, y es posible especificar la dirección de la escala según la categoría positiva.

## 10.4. GALERÍA DE GRÁFICOS INTERACTIVOS

Los gráficos interactivos son mucho más potentes que los gráficos normales. Permiten especificar su formato con multitud de opciones y su creación es mucho más gráfica e intuitiva. Sin embargo, como ya se ha comentado, los requerimientos de recursos (memoria, espacio en disco y procesador) son muy grandes; por lo que es conveniente usarlos con una cantidad reducida de datos; de lo contrario, el tiempo necesario para crearlos puede ser excesivo. Además, hay ciertos tipos de gráficos que no es posible generar en su forma interactiva, como los de Pareto o lo de control.

Los gráficos disponibles en su configuración inicial son barras, puntos, líneas, áreas, sectores, dispersados, diagramas de caja, barras de error, histogramas, diagramas de dispersión, etc.

Para acceder a la Galería, se ha de buscar dicha opción dentro de los temas de la ayuda. Después, seleccione uno de los tipos y aparecerán todos los subtipos de gráficos posibles. Por último, pinchando en uno de los subtipos, aparece otra ventana de ayuda que austeramente indica cómo hacerlo.

### 10.4.1. CREACIÓN DE GRÁFICOS INTERACTIVOS

Cuando se accede a un gráfico de este tipo a través del menú Gráficos > Interactivos, se muestra un cuadro de diálogo distinto a los anteriormente revisados. Este nuevo cuadro posee varias pestañas:

Estas pestañas permiten acceder a las distintas configuraciones del gráfico. En la parte inferior se muestran los típicos botones de los cuadros de diálogo para aceptar o cancelar el procedimiento, etc.

Los cuadros del resto de gráficos son similares. La ***primera pestaña*** es la de selección de variables. Permite asignar variables a los ejes, designar agrupaciones, cambiar a 3D (si es posible), aplicar funciones para representar las variables, etc. Para asignar variables, hay que arrastrarlas con el mouse hasta la casilla correspondiente. Para quitarla, arrástrela de nuevo a su sitio o bien sitúese otra en su lugar. Con la tecla Ctrl, se puede arrastrar varias.

La lista de origen (en la parte izquierda) presenta una relación de las variables disponibles para el tipo de gráfico seleccionado. Se puede cambiar el orden de las variables en la lista con el menú contextual. También, se puede elegir entre mostrar las etiquetas de las variables o sus nombres.

El tipo de variable se indica mediante un símbolo a su izquierda: Función de escala usada para el eje dependiente, resume casos o variables. Variable de escala: variable categórica.

La ***segunda pestaña*** es la de configuración del gráfico en sí. Dependiendo del tipo que sea aparecerán unas opciones u otras. En el caso de ejemplo, que es un gráfico de barras, se mostrarán opciones para cambiar la forma de las barras, las etiquetas del gráfico, etc.

El gráfico puede tener más pestañas de configuración propia del gráfico o de barras de error (si son posibles), en cuyo caso también se podrá modificar el formato de estas barras.

La ***tercera pestaña*** ofrece tres paneles para introducir el título, subtítulo y el pie del gráfico.

La ***cuarta pestaña*** es la de opciones generales (independientes del tipo de gráfico), como los órdenes categóricos, las escalas, la longitud de los ejes, el aspecto, etc.

### 10.4.2. GRÁFICOS DE BARRAS INTERACTIVOS

Presenta estadísticos de resumen de una o más variables, la mayor parte de las veces respecto a los grupos definidos por una o dos variables categóricas. La longitud de las barras representa habitualmente la frecuencia de casos de cada categoría, un porcentaje del número total de casos o una función de otra variable (por ejemplo, el valor medio para cada categoría). Las posibilidades gráficas son muy extensas.



#### A. Pestaña Asignar variables

La función se puede cambiar en la parte inferior del cuadro de diálogo y mostrarse una clave para identificar la función. Permite cambiar la orientación, de vertical a horizontal, y la representación en 3D para añadir otra variable, así como simplemente generar un efecto 3D. Si se asigna una variable categórica a Color o Estilo, las barras de las categorías se podrán agrupar unas junto a otras, o apilarse. En las barras apiladas, las categorías de la variable de la leyenda se apilan unas sobre otras. La parte superior de un segmento conforma la base del siguiente.

Este tipo de gráficos destaca la suma de las categorías; por tanto, solo se deberán seleccionar funciones adecuadas para la suma en *Las barras representan*. Si se ha seleccionado una leyenda de apilado de color o de estilo, se podrá cambiar el eje dependiente para que represente el 100%. En ese caso, los segmentos representarán el porcentaje con el que cada categoría contribuye al total.

Si se asigna la variable preincorporada *CASO* a un eje, aparecerá otra opción disponible en la que se podrá asignar una variable para etiquetar los casos. Por ejemplo, las etiquetas pueden ser días de la semana o meses.

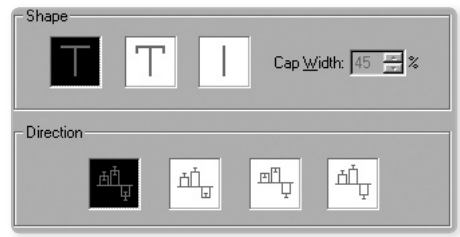
#### B. Pestaña Opciones del gráfico de barras

Existen varias opciones para mejorar los gráficos de barras, como modificar su forma, etiquetar las barras o establecer el valor de la línea base de las barras. En el ejemplo mostrado, se visualizan las formas posibles de las barras y la selección de base, cuadrada o circular.



### C. Pestaña Barras de error

Las barras de error, que indican la variabilidad de la medida mostrada, se encuentran disponibles cuando la variable dependiente es una variable de escala y las barras o las líneas representan la función de resumen *Medias*. La longitud de las barras de error se muestra como los límites inferior y superior de un intervalo de confianza (para un nivel porcentual especificado), o en desviaciones típicas o errores típicos de la media. Se puede especificar el nivel del intervalo de confianza, la dirección y la forma de las barras de error y cambiar su aspecto. En el ejemplo, se observan las diferentes formas que pueden tomar los "bigotes", el porcentaje de anchura del remate y la dirección de las barras:




### D. Pestaña Títulos

Se puede añadir un título y un subtítulo al gráfico, así como establecer un pie.

### E. Pestaña Opciones

Por último, en esta ficha se muestran otros métodos para personalizar los resultados. Se puede determinar el orden de las variables categóricas y el rango de las variables de escala.

También se puede elegir un *Aspecto de gráfico* y el tamaño de la región de datos. Las variables categóricas se pueden ordenar por una propiedad (valores, etiqueta, apariciones o frecuencia) o por una función de resumen de otra variable.

Además, se puede determinar que el orden sea ascendente o descendente y excluir categorías sin datos. Se puede establecer el valor mínimo y el máximo de todas las variables de escala, utilizar los valores predeterminados del conjunto de datos o crear una amplitud personalizada. La lista *Aspectos de gráfico* contiene varios aspectos almacenados en el directorio \Looks (por defecto). Estos aspectos controlan las propiedades visuales iniciales del gráfico, tales como los colores y los símbolos.

El *Tamaño* de la región de datos viene determinada por la longitud de cada eje. Si se asigna una variable de panel, la longitud de los ejes Y y X1, se utilizará para determinar la relación de aspectos de los paneles.

## 10.4.3. GRÁFICOS DE PUNTOS INTERACTIVOS

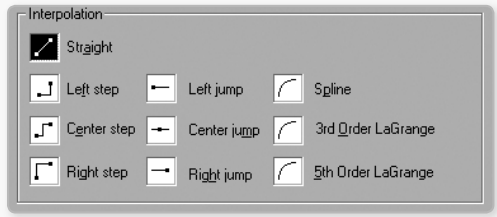
Su objeto es el mismo que el de los gráficos de barras. Resumen categorías de una o más variables. Representa puntos en los puntos de datos, pero también líneas que conectan los puntos o tanto líneas como puntos. Ejemplos:



Las pestañas son las mismas que en el caso anterior, a excepción de las opciones propias del gráfico que puede usar para modificar las especificaciones particulares de este tipo de gráfico. Esta pestaña se emplea también en líneas, bandas y líneas verticales.

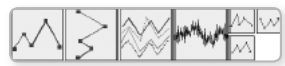
**Ejemplo:**

Una de estas opciones propias es la de interpolación, que controla el modo de dibujar las líneas entre los puntos de datos. Las elecciones posibles serían las siguientes:

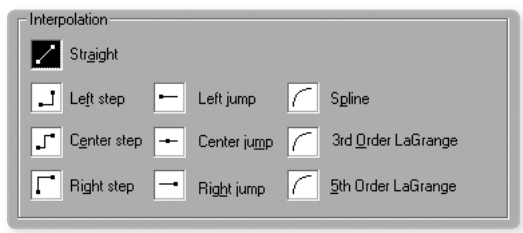



### 10.4.4. GRÁFICOS DE LÍNEAS INTERACTIVOS

Realmente, este tipo de gráfico funciona al unísono con el anterior y con el de líneas verticales. Solo se debe seleccionar en las opciones de *puntos y líneas* si se quiere mostrar puntos, líneas o líneas verticales. Ejemplos:



Las líneas, en concreto, tienden a enfatizar el flujo o los movimientos y no los valores individuales. Se suelen utilizar para representar datos a lo largo del tiempo, y por tanto, pueden usarse para observar tendencias. Los gráficos de líneas son especialmente eficaces para representar un gran número de valores en un espacio compacto.

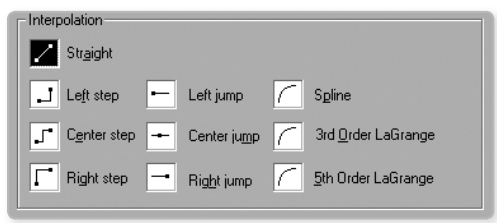



### 10.4.5. GRÁFICOS DE BANDAS INTERACTIVOS

Un gráfico de bandas es similar a uno de líneas, pero las líneas se representan como bandas en una tercera dimensión. Los gráficos de bandas pueden ser bidimensionales (con un efecto 3D) o tridimensionales. Ejemplo:



El resto es idéntico a los de líneas y puntos:

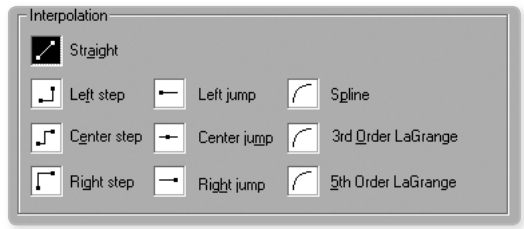


### 10.4.6. GRÁFICOS DE LÍNEAS VERTICALES INTERACTIVOS

En los gráficos de líneas múltiples o de puntos, en los que se asigna una variable de categoría a *Color* o *Estilo*, se pueden dibujar líneas verticales entre los puntos de color o de estilos del mismo eje de categoría. Ejemplo:



En cuanto al resto, se recurrirá a los gráficos de puntos y líneas:




### 10.4.7. GRÁFICOS DE ÁREAS INTERACTIVOS

En un gráfico de áreas simple, los puntos se representan y se unen con líneas, y el área comprendida entre la línea de datos y la línea base se rellena. Uno de los usos más comunes de estos gráficos consiste en mostrar la continuidad de una variable de resumen (dependiente) en el tiempo (variable independiente). Posibilidades:



Lo único que varía respecto a los anteriores tipos de gráficos son las opciones propias, que en este caso solo le permiten configurar las etiquetas y los valores perdidos.

### 10.4.8. GRÁFICOS DE SECTORES SIMPLES INTERACTIVOS

Resume las categorías definidas por una sola variable o por un grupo de variables relacionadas. En cada categoría, el tamaño de cada sector representa la frecuencia, el porcentaje o una función de resumen de una variable. También se puede crear un gráfico de sectores con paneles. Ejemplos:



Las opciones le permiten establecer las etiquetas y la posición. Por ejemplo, se ve cómo cambiar la dirección y el ángulo inicial:




El resto de opciones no difiere de las de los tipos de gráficos anteriores.

### 10.4.9. GRÁFICOS DE SECTORES AGRUPADOS INTERACTIVOS

Contiene una agrupación de gráficos de sectores simples, todos ellos apilados como categorías de una misma variable. Los gráficos poseen diferentes tamaños y aparecen apilados unos sobre otros. La agrupación contiene tantos gráficos de sectores, como categorías hay en la variable de agrupación. Vea el siguiente ejemplo:



Las opciones le permiten establecer las etiquetas y la posición. Por ejemplo, se verá cómo cambiar la dirección y el ángulo inicial.




### 10.4.10. GRÁFICOS DE SECTORES DISPERSADOS INTERACTIVOS

Contiene un conjunto de gráficos de sectores situados en puntos definidos, por coordenadas rectangulares. Es similar a un diagrama de dispersión con sectores como marcadores gráficos. Todos los gráficos están divididos en sectores por la misma variable categórica. Las opciones le permiten establecer las etiquetas y la posición. Por ejemplo, vea cómo cambiar la dirección y el ángulo inicial:




#### A. Diagramas de caja interactivos

Muestran la mediana, los cuartiles, los valores atípicos y los valores extremos para una variable de escala. Pueden dividirse en categorías y agrupaciones. Posibilidades:



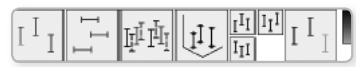
En las opciones se pueden mostrar los valores atípicos, los extremos y las líneas de la mediana de cada caja. Los valores atípicos se encuentran a una distancia del final de la caja de entre 1,5 y 3 veces la longitud de esta. Los extremos se encuentran a una distancia del final de la caja de más de tres veces la longitud de la caja. En los diagramas de caja 2D, se puede elegir entre varias formas de remate para cada bigote. En las cajas 3D, la forma de la base puede ser circular o cuadrada. Por último, si *Mostrar etiquetas de frecuencia* está seleccionado, mostrará el número de casos incluidos en cada categoría.

**B. Barras de error interactivas**

Las barras de error facilitan la comprobación en el gráfico de las distribuciones y la dispersión, indicando la variabilidad de la medida mostrada. La media de una variable de escala se representa para un conjunto de categorías; la longitud de una barra de error a cada lado del valor medio indica un intervalo de confianza, o un número especificado de errores típicos o de desviaciones típicas.

Las barras de error pueden extenderse en una o en ambas direcciones desde la media. A veces, en el mismo gráfico, se muestran las barras de error junto a otros elementos gráficos, como las barras.

Las opciones le permiten modificar las etiquetas, la dirección o la forma, así como si quiere ver una marca en el valor medio de cada barra de error. Ejemplos:



**C. Histogramas interactivos**

Reúne los valores de una variable en grupos espaciados regularmente (intervalos o clases) y representa el recuento del número de casos dentro de cada grupo. Este recuento o frecuencia se puede expresar en forma de porcentaje, el cual es especialmente útil para comparar conjuntos de datos de diferentes tamaños. La frecuencia o el porcentaje también pueden acumularse a lo largo de los grupos.

El histograma permite detectar los valores atípicos y las desviaciones de la simetría. Dichas desviaciones pueden indicar que una variable no es adecuada para ser analizada mediante un procedimiento que asuma una distribución normal; a través de la pestaña Histograma, se puede superponer una curva normal (solo disponible para 2D) con la misma media y varianza que las de los datos. Permite situar intervalos y punto de inicio de la variable sobre un eje determinado.

Puede establecer el número de intervalos (máximo 250) o su anchura; de lo contrario, se ajustará automáticamente. El punto inicial del primer intervalo puede establecerse, en cualquier lugar, desde el valor mínimo de escala hasta el valor mínimo, más la anchura de un intervalo. Ejemplos de distintas configuraciones:



**D. Diagramas de dispersión interactivos**

Estos diagramas destacan la relación existente entre dos o tres variables cuantitativas mediante la representación de los valores reales, a lo largo de dos o tres ejes. A menudo, es posible observar relaciones o un patrón curvilíneo; por ejemplo, que los estadísticos descriptivos no revelan.

También es posible descubrir valores atípicos bivariados. Asignando una variable de leyenda, se pueden estudiar las características de los subgrupos. Además, pueden utilizarse los valores de una variable para etiquetar los puntos; de lo contrario, se emplearán los números de caso como etiquetas.

Para ajustar líneas a los puntos de un diagrama de dispersión, se dispone de los métodos regresión, media, suavizador o ninguno. Con el método regresión, se puede indicar la inclusión de una constante en la ecuación. Con el suavizador, se puede indicar el tipo de kernel, los multiplicadores del ancho de banda, etc.

Con los métodos media o de regresión, se pueden mostrar las líneas del pronóstico promedio, o bien las líneas del pronóstico individual para ilustrar el nivel de confianza especificado. Además de una línea de ajuste para la nube total, si existe una división categórica, se pueden seleccionar las líneas de ajuste para los subgrupos.

Los **trazos de unión** son líneas dibujadas entre una nube de símbolos y punto, eje, línea o superficie. Estos trazos pueden resultar útiles para leer los valores de los ejes o para comparar distancias mediante la longitud de las líneas. Los trazos que se dibujan hasta un eje horizontal se denominan también «líneas verticales». Algunos ejemplos son los siguientes:



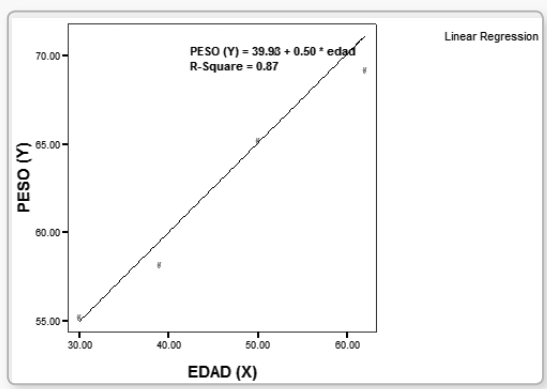


Además de la fórmula de la recta de regresión, resulta necesario disponer de información sobre el grado en que el modelo se ajusta a los datos observados (nube de puntos).

## Ejemplo: Bondad de ajuste del modelo

Además de la fórmula de la recta de regresión, resulta necesario disponer de información sobre el grado en que el modelo se ajusta a los datos observados (nube de puntos).

Una primera aproximación es la interpretación gráfica del problema. Para elaborar la gráfica del ajuste de la recta a los datos observados, se pulsa el menú Gráficos > Interactivos > Diagramas de dispersión. Luego, el cuadro de diálogo de Diagramas de dispersión interactivos es el siguiente: Defina la variable del criterio (en el eje de ordenadas) y la variable predictora (en el eje de abscisas), desde la solapa Asignar variables. A continuación, se selecciona el método Regresión desde la solapa Ajuste (Fit). El cuadro de diálogo desde el que se hacen estas selecciones tiene el siguiente aspecto:




Como se observa, los puntos no se alejan mucho de la recta; luego el ajuste es aceptable. Además de la interpretación gráfica del problema, la forma de cuantificar la bondad del modelo es mediante el coeficiente de determinación: $r^2_{XY}$.

Se trata de una medida estandarizada que toma valores entre 0 y 1 y cuya interpretación es muy sencilla: representa la proporción de varianza explicada de la variable del criterio, a partir de la predictora. En este ejemplo, $r^2_{XY}$ = 0,875; por lo que el modelo sí es adecuado para hacer pronósticos de Y, a partir de X.

## 10.5. CREACIÓN DE GRÁFICOS SEGÚN LA ESTRUCTURA DE LOS DATOS

Para la creación de un gráfico, es fundamental tener claro lo que se desea obtener y el tipo de gráfico adecuado a su estructura de datos. A partir de aquí, se establece una clasificación de los gráficos que podría utilizar, según las variables de las que consten sus datos.

### A. Gráficos para variables categóricas

- **Para una única variable categórica.** Se puede utilizar barras simples, líneas simples, áreas simples, sectores y Pareto simple. Siempre se deben seleccionar resúmenes para grupos de casos, ya que se trata con una variable única.
- **Categorías de una variable dentro de categorías de otra variable.** Barras agrupadas o apiladas, líneas múltiples o verticales, áreas apiladas, Pareto apilados, máximo-mínimo-cierre simples, barras de rango simples y líneas de diferencia. El análisis para todos ellos es por grupos de casos.

### B. Gráficos para variables continuas

- **Variable continua única.** Resúmenes para distintas variables en diagramas de caja simples o en barras de error simples, histogramas, P-P, Q-Q; los casos son subgrupos en p, np y en gráficos c,u, secuencias, autocorrelaciones y espectral.
- **Variable continua única dentro de categorías de otra variable.** Resúmenes para distintas variables en diagramas de caja simples o en barras de error simples, los casos son unidades en X-barra, R, s; p, np y c,u.
- **Varias variables continuas.** Resúmenes para distintas variables en barras simples, líneas simples, sectores, diagramas de caja simples, barras de error simples; diagrama de dispersión simple, superpuesto matricial y 3D; sumas para distintas variables en Pareto simples, los casos son subgrupos en X-barra, R, s; secuencia, autocorrelaciones, correlaciones cruzadas y espectral.
- **Varias variables continuas dentro de categorías de otra variable.** Resúmenes para distintas variables en barras agrupadas o apiladas, líneas múltiples y verticales, áreas apiladas, diagramas de caja agrupados, barras de error agrupadas; diagrama de dispersión simple con variable de control, Pareto apiladas, máximo-mínimo-cierre simples, barras de rango simples, líneas de diferencia simples; resúmenes para grupos de casos en máximo-mínimo-cierre agrupados y en barras de rango agrupadas.

### C. Gráficos para casos individuales

- **Casos individuales (variable única).** Valores de los casos individuales en barras simples, líneas simples, áreas simples, sectores, Pareto simples y los casos son unidades en Individuo, rango móvil.
- **Casos individuales (dos o más variables).** Barras agrupadas y apiladas, líneas múltiples y verticales, áreas y Pareto apilados, máximo-mínimo-cierre simples y agrupados, barras de rango simples y agrupadas y líneas de diferencias. Siempre con la opción Valores de los casos individuales.

### D. Gráficos especializados

- **Variable única de máximo-mínimo-cierre dentro de categorías.** Resúmenes para grupos de casos en máximo-mínimo-cierre simples, barras de rango simples y líneas de diferencia.

- **Variables distintas de máximos y mínimos dentro de categorías.** Resúmenes para distintas variables en máximo-mínimo-cierre simple y agrupado, barras de rango simples y agrupadas, líneas de diferencia. Resúmenes para grupos de casos en máximo-mínimo-cierre agrupados y barras de rango agrupadas.
- **Distintas variables en máximos y mínimos por caso individual.** Valores de los casos individuales en máximo-mínimo-cierre simples y agrupados, barras de rango simples y agrupadas, líneas de diferencia.
- **Control de procesos. Variable de medida única.** Los casos son unidades en Individuo, rango móvil. Los casos son subgrupos en p,np; c,u.
- **Control de procesos. Variable única dentro de categorías.** Los casos son unidades en X-barra, R, s; p, np; c,u. Los casos son subgrupos en X-barra, R, s.
- **Series temporales.** Secuencia, autocorrelaciones, correlaciones cruzadas y espectrales.

CAPÍTULO

11

# PERSONALIZACIÓN SPSS

Se aprenderá a personalizar el software SPSS de acuerdo a los requerimientos y necesidades, utilizando además las principales opciones -posibilitando así la configuración de SPSS a nivel general-, que se encuentran distribuidas en pestañas de acceso a diferentes fichas. Asimismo, se podrá personalizar la barra de herramientas, dependiendo del tipo de documento que seleccione; será susceptible de mostrar, ocultar, modificar, crear una nueva barra asignándole un nombre y posteriormente personalizarla. Además, se podrá modificar la barra de menú, insertar un nuevo menú o un separador en cualquier posición de la lista; es decir, se podrá crear un menú dentro de otro menú. Por último, tendrá una noción de procesamiento y autoprocesamiento.

## 11.1. CUADRO GENERAL DE OPCIONES

Este cuadro se encuentra accesible desde cualquier ventana o editor, a través del menú Edición > Opciones.




Posibilita la configuración de SPSS a nivel general y se encuentra distribuido en pestañas de acceso a diferentes fichas. Las modificaciones que se realizan aquí, por norma general causarán efecto cuando el sistema deba volver a utilizarlas.

### 11.1.1. DESCRIPCIÓN DE LAS OPCIONES

#### A. Ficha General

Luego de pulsar en opciones, aparece el cuadro de diálogo de la ficha General, tal como se observa en la figura.

El *Diario de la sesión* registra todos los comandos ejecutados en una sesión, tanto en ventanas de sintaxis como en cuadros de diálogo. Este archivo se puede editar y reutilizar. El *sistema de medida* utilizado puede basarse en puntos, pulgadas o centímetros. Los usuarios avanzados que usen comandos usualmente pueden establecer que se abra una ventana de sintaxis automáticamente al inicio de cada sesión. Las *listas de variables* controlan la presentación de las mismas y su orden en los cuadros de diálogo. El *tipo de resultados al arrancar* controla el tipo de visor que se abre en primera instancia. La *notificación de resultados* es la forma de indicar al usuario que el procedimiento ha terminado y los resultados están listos en el visor.

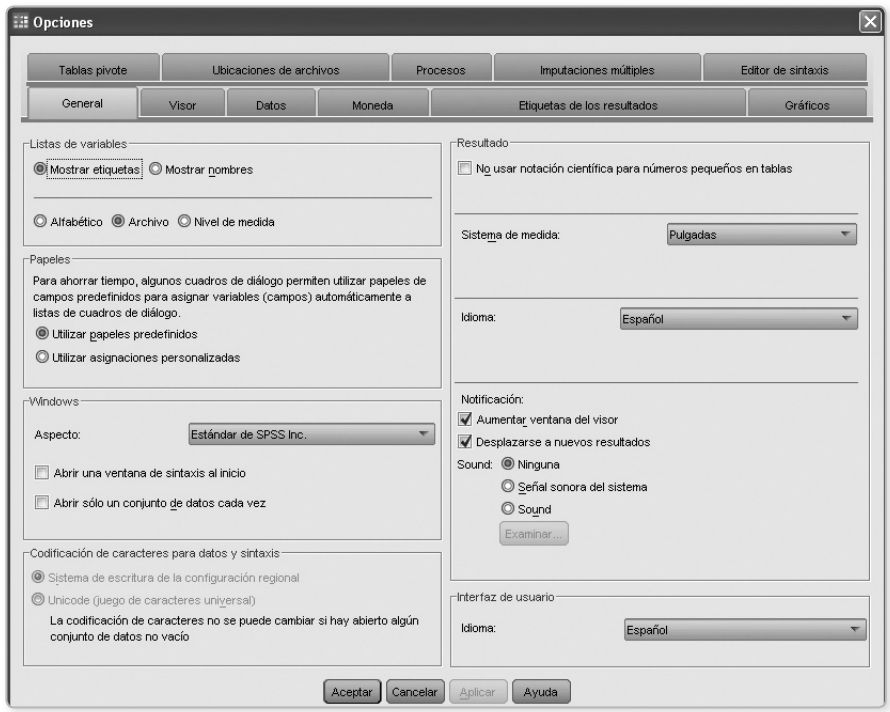



*Figura 11.1 Cuadro de diálogo de la ficha General*

**B. Ficha Visor**

Controla los elementos que se muestran y se ocultan automáticamente cada vez que se ejecuta un procedimiento, además de la alineación inicial de los elementos, pero solo en la impresión, no en pantalla.

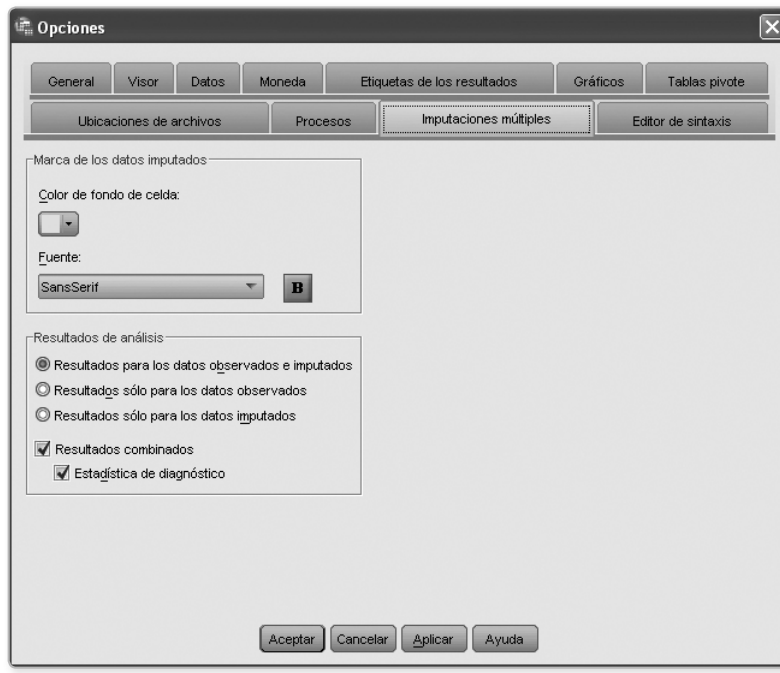



*Figura 11.2 La fuente del título cambia el estilo, tamaño y color.*

El *tamaño de página de los resultados de texto* controla el ancho y largo de la página expresada en caracteres y líneas, respectivamente. La *fuente de los resultados de texto* también se puede modificar, pero usando siempre las que posean un ancho de letra homogéneo.

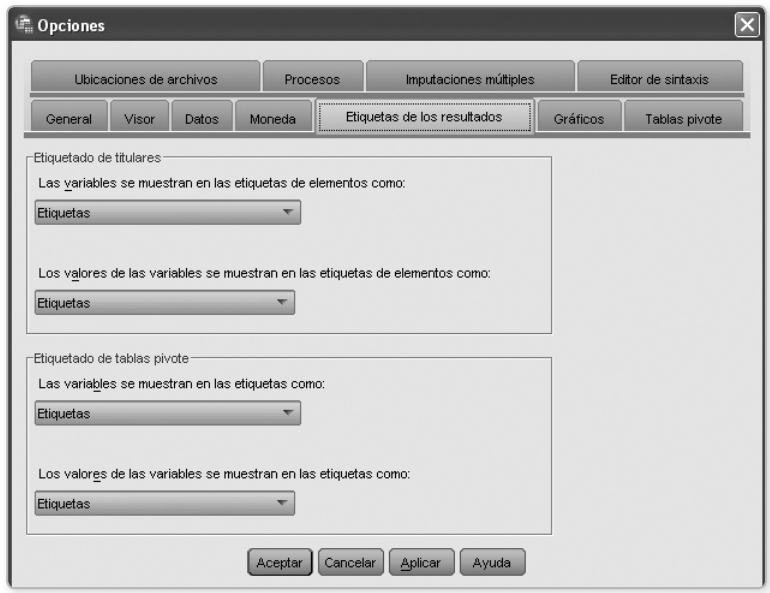
### C. Imputaciones múltiples

- **Mostrar elementos de resultados.** Determina los elementos que deben mostrarse automáticamente cada vez que se ejecute un procedimiento. Se pueden insertar saltos de página entre resultados, así como cambiar la fuente por defecto.
- **Resultados tabulares.** Controla las opciones de los resultados de las tablas pivote convertidos a resultados tabulares.
- **Resultados de texto.** Que no sean de tablas pivote convertido, se puede establecer el ancho y la longitud de página.



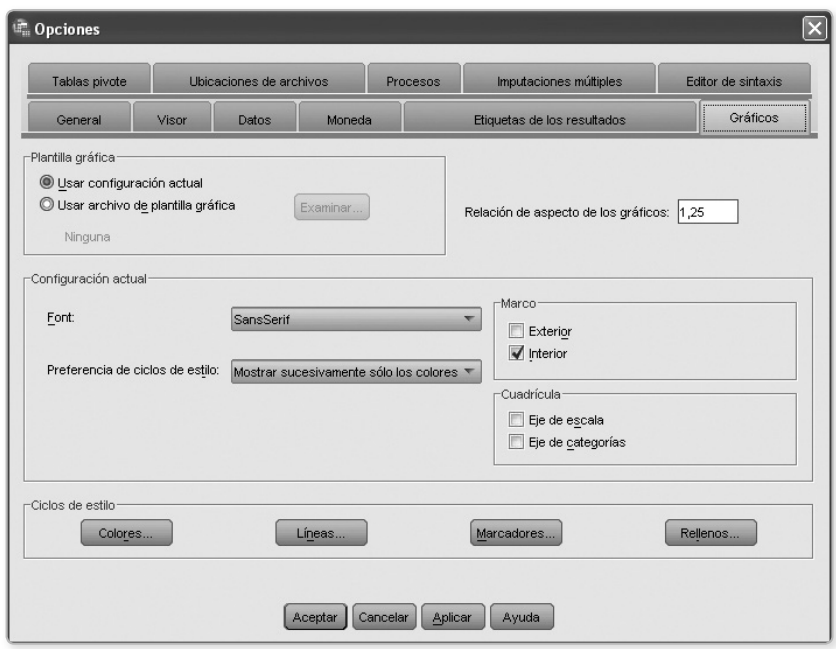

### D. Ficha Etiquetas de los resultados

Controlan la visualización de la información sobre variables y valores de datos en las tablas pivote y en los titulares. Puede visualizar nombres de variable, etiquetas de variable definidas y valores de datos reales, etiquetas de valor definidas o una combinación de estas.



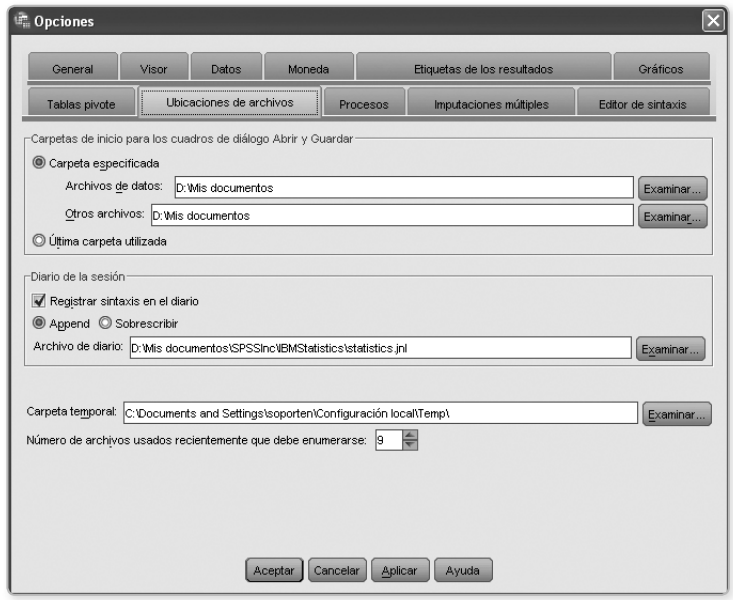

### E. Ficha Gráficos

Los archivos nuevos pueden utilizar las opciones establecidas en *Especificaciones actuales,* o las de una plantilla. La *relación de aspecto de los gráficos* indica la proporción ancho-alto del marco exterior de los nuevos gráficos. Las *especificaciones actuales* controlan la fuente de los textos en los gráficos, las tramas de relleno y estilos de línea, los marcos internos y externos, y la visualización de las líneas de cuadrícula. Todo esto no afecta a los gráficos interactivos.



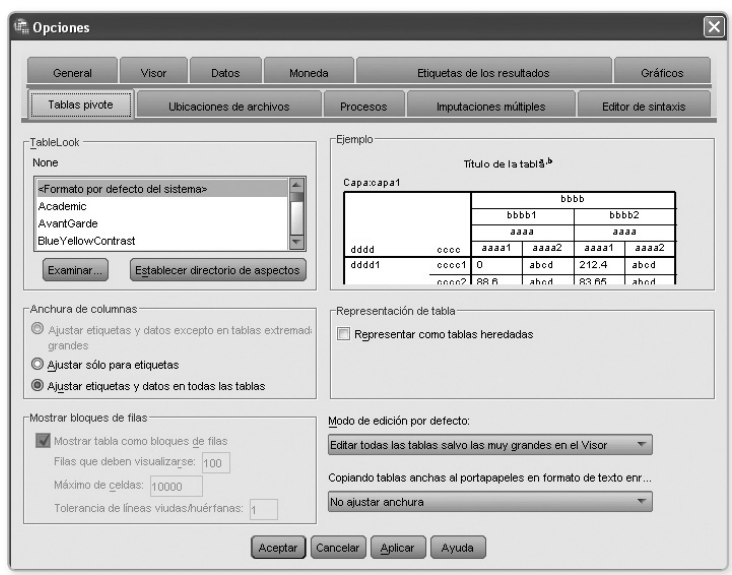
### F. Ubicación de archivos

Permite establecer el *aspecto del gráfico* inicial y las *unidades de medida* usadas. En el apartado de gráficos interactivos ya se vieron las ventajas e inconvenientes de guardar datos con el gráfico. También es posible determinar la *resolución de impresión* de estos gráficos.




### G. Ficha Tablas pivote

Igual que en los gráficos, permite establecer un *aspecto de tabla* por defecto. Se puede ajustar automáticamente el ancho de las columnas en las tablas pivote. El *modo de edición por defecto* establece la forma de edición de las tablas, bien en una nueva ventana de edición, bien en la ventana del visor donde se encuentra.



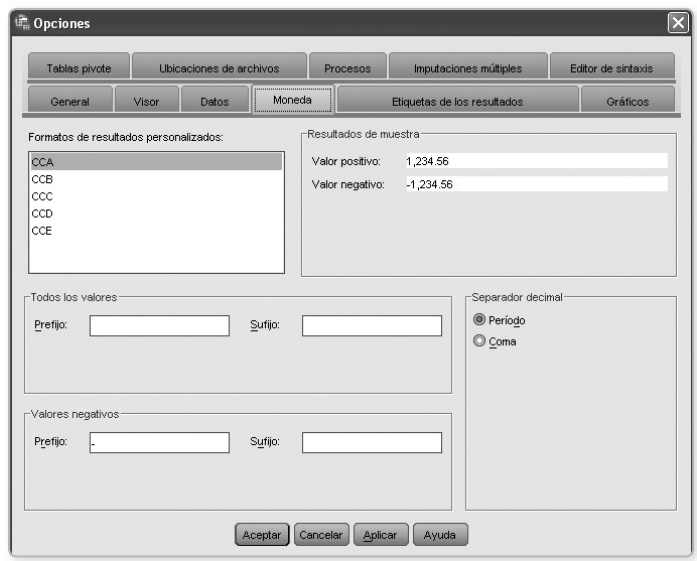
### H. Ficha Datos

Ya se vio qué ventajas aportaban las *opciones de transformación y fusión* en el apartado que describía el menú Transformar. El formato de *visualización para nuevas variables numéricas* controla la visualización por defecto del ancho y el número de posiciones decimales de las nuevas variables numéricas. Por defecto, no hay formato para las de cadena. *Definir rango de siglo para años de dos dígitos* hace referencia a las variables con formato de fecha introducida o mostrada con un año de dos dígitos. Es una buena solución para mantener dos dígitos aparentes en las fechas tras el pánico del Y2K.



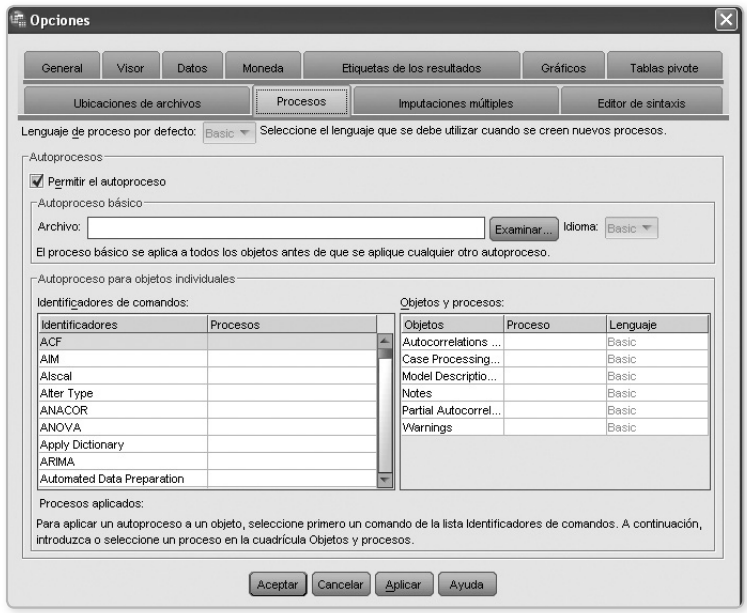

### I. Ficha Moneda

Pueden crearse hasta cinco formatos de visualización de moneda personalizados, que pueden incluir caracteres de prefijo y sufijo especiales, tratamiento especial para los *valores negativos* e indicación del *separador decimal* a usar.



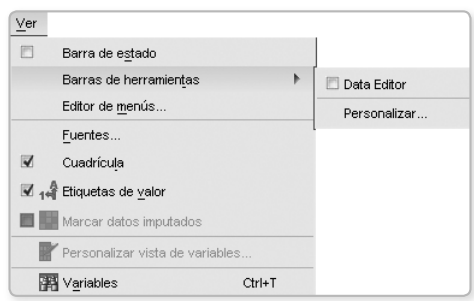
### J. Ficha Procesos

Especifica el *archivo de procedimientos global*, el *archivo de autoproceso* y las subrutinas de autoproceso a usar. Un *archivo de procedimientos globales* es una biblioteca de subrutinas de proceso y funciones que pueden ser invocadas por archivos de proceso, incluyendo los archivos de autoproceso. Estos términos se explican en el apartado de *Procesamiento*. En cuanto al resto, se concreta en el apartado *Autoprocesamiento*.




## 11.2. PERSONALIZACIÓN DE LAS BARRAS DE HERRAMIENTAS

Para personalizar las barras de botones de cualquier ventana, acceda a Ver > Barras de herramientas.




Dependiendo del tipo de documento que seleccione, se mostrarán las barras asociadas a dicha ventana que son susceptibles de mostrar u ocultar. Se puede crear una nueva barra asignándole un nombre y posteriormente personalizándola; o bien, puede elegir una de las barras que posee SPSS y modificarla.

También puede cambiar el tamaño de los botones (cuadruplicando su tamaño) o eliminar las pistas que le ofrecen cuando detiene el puntero del mouse unos segundos sobre cualquiera de ellos.

Una vez en el cuadro Personalizar, elija la categoría Global de herramientas. En la lista de la derecha, se mostrarán los botones asociados a esa categoría y únicamente debe arrastrarlos a la barra modelo que aparece en la parte inferior para que se inserte.

Con Editar herramienta, se abre una nueva ventana capaz de editar un pequeño ícono de tipo BMP de 16x16 puntos con 16 colores (tamaño de un botón estándar). Una vez acabado el diseño, seleccione Aplicar y salir, con lo que ya tendrá el aspecto del botón a su gusto.

La opción Propiedades permite asignar la barra que esté editando a las ventanas que quiera. De esta forma, cuando se encuentre en una de esas ventanas, dicha barra se mostrará por defecto.

Si desea crear una nueva herramienta, dele un nombre (etiqueta) e indique la acción a realizar, que puede ser ejecutar una aplicación, un archivo de sintaxis o un proceso. Por ejemplo, seleccione Aplicación, indique los datos de tipo XLS y en el nombre de archivo coloque Excel.exe. Le aparecerá una nueva herramienta con el nombre que le haya dado en la categoría Personales (que es donde se guardan las creadas por el usuario), y ya solo queda arrastrar dicho botón a la barra que esté editando.

Siempre puede restablecer la barra original si la barra creada no se ajusta a sus necesidades o ya le ha dado el uso que pretendía.

## 11.3. PERSONALIZACIÓN DE LAS BARRAS DE MENÚS

Para crear elementos de menú personalizados, seleccione el Editor de menús en el menú Utilidades.




Ahora, elija la barra de menús que quiera modificar y en el panel de la izquierda aparecerá una lista jerárquica con todas las opciones de dicha barra. Seleccione la posición dónde insertar el nuevo elemento y defínalo. Este elemento ejecuta una aplicación, una sintaxis o un proceso, de la misma forma que se vio al crear un nuevo botón en la barra de herramientas. Asimismo, puede insertar un nuevo menú o un separador en cualquier posición de la lista; es decir, se podría crear un menú dentro de otro menú. El símbolo **&** indica que el carácter siguiente aparecerá subrayado y generará un acceso rápido mediante la combinación de teclas Alt y dicho carácter.

## 11.4. PROCESAMIENTO

El software SPSS dispone de una utilidad de procesamiento para automatizar tareas que se realicen con frecuencia. Tareas como personalización de resultados, apertura y almacenamiento de archivos, trabajo con cuadros de diálogo, transformación de datos, procedimientos estadísticos, exportación de gráficos, etc.

El sistema incluye una serie de procesos de muestra, que pueden servir para iniciarse en su creación y edición. Algunos ejemplos son los siguientes:

- **Limpiar navegador**. Elimina las tablas Notas de un documento de resultados.
- **Totales en negrita**. Aplica negrita y azul a cualquier fila o columna con la etiqueta Total en una tabla pivote.
- **Borrar por etiqueta**. Elimina filas o columnas de una tabla pivote basándose en el contenido de las etiquetas de fila o columna.

Los procesos se ejecutan a través del menú Utilidades. Dependiendo de la función que realice el proceso, es usual que se deba seleccionar previamente el objeto al que se quiera aplicar dicha función (por ejemplo una tabla pivote).

Para crear un proceso, diríjase al menú Archivo > Nuevo > Proceso. Se muestra una serie de procesos de inicio (directorio \Starter) que se pueden usar, o bien pulse Cancelar e inicie de cero. Los procesos tienen la extensión **.sbs* y suelen contener una breve descripción en el panel de la derecha.

La ventana que aparece es el Editor de procesos, y se encontrará en un entorno de programación completo que utiliza el lenguaje Sax Basic, compatible con Visual Basic for Applications.

Incluye un Editor de cuadros de diálogo, un visor de objetos, funciones de depuración y ayuda sensible al contexto. Los usuarios habituados a la programación pueden encontrar útil la definición de este lenguaje ubicada en la ayuda del editor.

## 11.5. AUTOPROCESAMIENTO

El concepto de autoproceso es similar al de proceso, con una ligera diferencia: se ejecuta automáticamente siempre que se produzca determinado tipo de objeto de resultado.

### Ejemplo

Definir un autoproceso que aumente en dos puntos el tamaño de fuente de la primera fila de una tabla pivote.

Cada vez que se genere una tabla pivote, automáticamente se ejecutará el autoproceso y la modificará. Para usar los autoprocesos, diríjase al menú de opciones generales (menú Edición) y a la pestaña Procesos. Existe una casilla denominada Permitir el autoprocesamiento, la misma que se activará.

Elija el archivo de autoproceso y se mostrarán en la lista inferior las subrutinas que contiene (cada archivo puede contener varias). Solo resta activar las subrutinas que se quiere para que actúen automáticamente.

Para crear o editar un autoproceso, sitúese en el Visor y seleccione el objeto al que aplicará el autoprocesamiento, que actuará como desencadenante del mismo. En el menú Utilidades, se tiene la opción Crear > Editar autoproceso (también en el menú contextual), que le abrirá el Editor de procesos para añadir subrutinas al archivo actual de autoprocesos.

CAPÍTULO

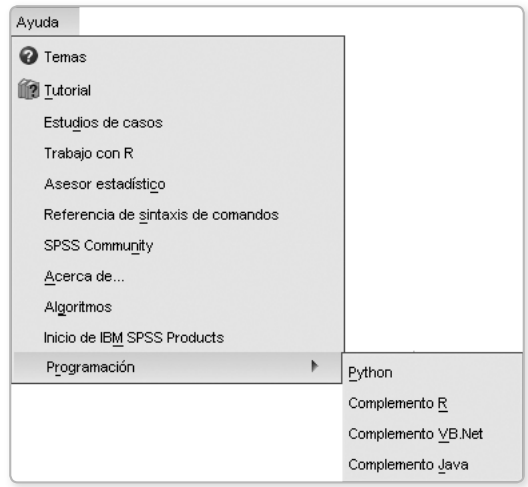
12

# AYUDA

La expresión "?" se refiere a la ayuda que ofrece el programa, que en todo momento será útil para el usuario. Puede accederse a él desde la ventana Editor de datos y Visor de. Desde Temas, puede escribirse un tipo de análisis (por ejemplo: Descriptivas) y el SPSS muestra una descripción de sus opciones. También, es posible obtener ayuda pulsando el botón derecho del mouse. Por ejemplo, si se encuentra en el cuadro de diálogo donde se calculan los estadísticos descriptivos y se quiere saber la definición de *media*, solo se necesita seleccionarlo y pulsar el botón derecho del mouse. También, puede obtener ayuda referente a guía de sintaxis, ayuda contextual, asesor de resultados, asesor estadístico.

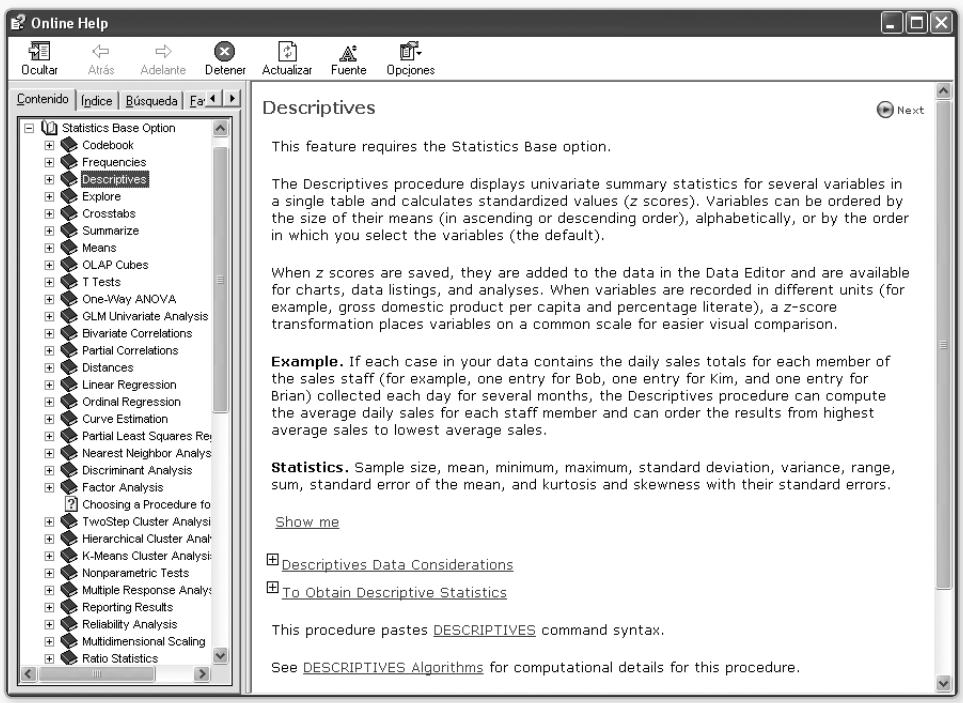
## 12.1. AYUDA

Desde Temas, puede escribirse un tipo de análisis (por ejemplo: Frecuencias) y el SPSS muestra una descripción de sus opciones. También es posible obtener ayuda pulsando el botón derecho del mouse.




**Ejemplo:**

Si se encuentra en el cuadro de diálogo donde se calculan los estadísticos descriptivos y se quiere saber la definición de curtosis, solo seleccione y pulse el botón derecho del mouse.



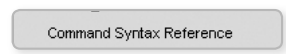
## 12.2. TEMAS

Ayuda de SPSS se encuentra en el menú ? donde se muestra un panel con fichas de *Contenido*, *Índice* y *Buscar*.

- **Contenido.** Se halla distribuido en libros que, al abrirlos, muestran más libros o temas de ayuda a exponer.
- **Índice.** Contiene una lista por la cual se indexa la ayuda. Introduciendo las primeras letras del término que se busca, irán apareciendo los objetivos de ayuda en la lista inferior.
- **Buscar.** Genera una base de datos con todas las palabras existentes en la ayuda, de forma que se puedan encontrar los términos más recónditos. Cabe resaltar que para maximizar las capacidades de búsqueda, se requiere recopilar una ingente cantidad de datos que puedan tener un elevado coste temporal. Cualquier tema que se visualice es susceptible de ser copiado, transportado a otra aplicación, impreso, etc. Uno de los libros de la ayuda es Tutoriales, donde la explicación progresiva de un tema expuesto se hace de una forma simple y didáctica.

## 12.3. GUÍA DE SINTAXIS

SPSS funciona internamente sobre la base de comandos.

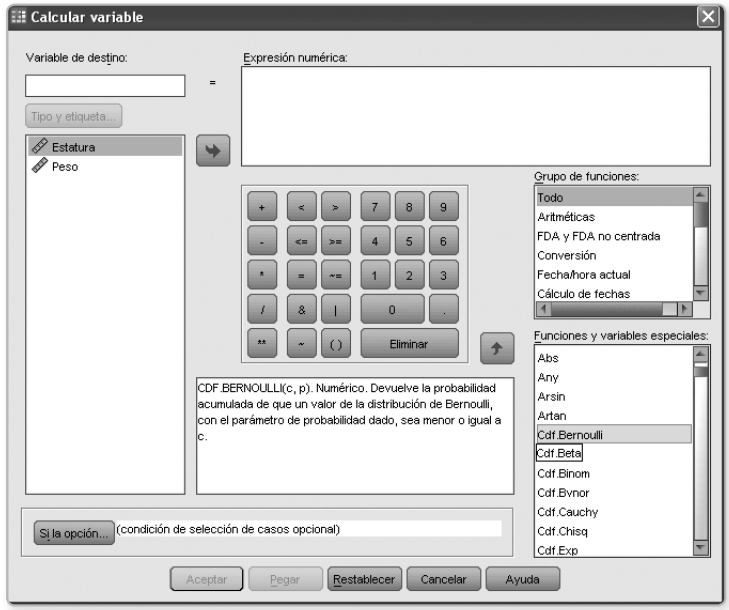



La mayoría de funciones se encuentran desarrolladas mediante la interfaz de ventanas, pero existen algunas que solo pueden ser implementadas mediante comandos.

La descripción de la sintaxis de comandos se encuentra en ficheros con formato PDF de Adobe Acrobat. Para poder verlos, se requiere de un visor de este formato. Los ficheros se distribuyen por módulos del sistema: *Base*, *Regresión Models*, *Categories*, etc. Por ejemplo, el módulo de modelos de regresión.

## 12.4. AYUDA CONTEXTUAL

La ayuda contextual se puede obtener de varias formas, según donde se encuentre. En los cuadros de diálogo, haga clic con el botón derecho del mouse sobre cualquier elemento para obtener una descripción del mismo. Dependiendo del elemento que se trate, se muestra un pequeño cuadro con información sobre el mismo o un menú contextual con la opción ¿Qué es esto?




En las tablas pivote, se debe obtener el menú contextual sobre la cabecera de una columna o fila, pulsando sobre la opción ¿Qué es esto? para que muestre una descripción de los valores contenidos en esa columna o fila. Por último, todos los cuadros de diálogo contienen un botón de ayuda.

## 12.5. ASESOR DE RESULTADOS

Se encarga de presentar ayuda sobre el tipo de objeto elegido.

**Ejemplo:**

Si tiene una tabla pivote en el Visor de resultados, edítela y muestre la barra de herramientas; así, un botón aparecerá y presentará una ventana explicativa a modo de tutorial del tipo de resultados que le ofrece la tabla.

### 12.5.1. ASESOR ESTADÍSTICO

Básicamente, le pregunta qué quiere hacer y le dice cómo hacerlo. La primera pantalla que se presenta es para solicitarle el objetivo que se persigue como, por ejemplo, crear informes, resumir datos, identificar relaciones, etc.

Seleccionando una de estas opciones se mostrará un ejemplo de lo que se obtendrá, y pulsando en Más ejemplos se presentarán otros resultados posibles. Con el botón Ayuda, la descripción de las opciones es más completa. Cuando tenga claro lo que quiere obtener, pulse *Siguiente*.

Luego le preguntará el tipo de datos de los que dispone, que pueden ser ordinales, categóricos, numéricos continuos, etc. o su estructura. Según la opción elegida en primera instancia, podrá solicitarse alguna información más, como el tipo de representación que desea, tipo de resumen, tipo de gráfico, etc. Una vez contestadas todas las preguntas que, como puede observar no tienen un nivel técnico, pulse Finalizar y se presentará un resumen que le ayudará a obtener los resultados deseados. Además, abrirá automáticamente el cuadro de diálogo correspondiente al procedimiento que debe emplear para obtener dichos resultados.

## 12.6. PRODUCCIÓN AUTOMATIZADA

El software SPSS posee la capacidad de funcionar de forma totalmente autónoma y automatizada. Para realizar un trabajo de producción automatizado, elabore un archivo de sintaxis de comandos (es un archivo con formato de texto). Se dispone de un Editor de sintaxis que se abre cuando se pulsa Archivo > Nuevo > Sintaxis, o puede usar el botón Pegar de cualquier cuadro de diálogo. Estos archivos poseen la extensión **.sps*.

**Ejemplo:**

Un archivo de sintaxis de comandos: Este archivo realiza estadísticas descriptivas, como la media desviación estándar, valor mínimo y máximo, de la variable Estado civil. Una vez que dispone del archivo de sintaxis, arranque la Utilidad de producción desde Windows, que se encuentra como un módulo aparte del programa. Solo es necesario añadir archivos de sintaxis al trabajo de producción y pulsar el botón Ejecución. Desde este mismo cuadro, decida si quiere que se muestren los resultados en el visor normal o en el de borrador, o bien que se impriman al terminar. Los archivos de trabajos de producción se pueden guardar y abrir desde el disco. Poseen extensión **.spp*.

## GLOSARIO DE TÉRMINOS

**Ajuste (Adjustment).** Conjunto de técnicas que se emplean después de la recogida de datos, para controlar o tener en cuenta el efecto de las variables de confusión, sean conocidas o potenciales (sinónimos: controlar, tener en cuenta, estandarizar).

**Aleatorio (Randomized).** Al azar, a la suerte.

**Análisis (Analysis).** Comparación del desenlace del grupo de estudio, con el del grupo de control o testigo.

**Apareamiento (Matching).** Análisis conjunto de dos o más observaciones realizadas en el mismo individuo o en individuos similares. El apareamiento por dúos (o pares) es un tipo especial de apareamiento, en el que dos observaciones se analizan conjuntamente.

**Apareamiento por grupos (Group matching).** Método de apareamiento empleado durante el proceso de asignación en una investigación, en la cual los individuos del grupo de estudio y del grupo de control se seleccionan, de tal forma que la distribución de cierta variable o variables sea prácticamente idéntica en ambos (sinónimo: apareamiento por frecuencias).

**Apareamiento por pares (Pairing).** Forma especial de agrupación en la cual cada individuo del grupo de estudio se empareja con uno del grupo de control y se comparan sus desenlaces. Cuando se emplea este tipo de agrupación, se deben utilizar pruebas estadísticas especiales. Estas técnicas pueden aumentar la potencia estadística del estudio.

**Asignación (Assignment).** Selección de individuos para los grupos de estudio y de control.

**Asignación a ciegas (Blind assignment).** Proceso por el que los individuos son asignados al grupo de estudio o al de control, sin que ni ellos ni el investigador sepan a qué grupo se asignan. Cuando los sujetos y el investigador están "cegados", el estudio a veces se denomina «estudio doble ciego».

**Asignación al azar (Randomization).** Método de asignación en el cual los individuos tienen una probabilidad conocida; aunque no necesariamente igual, de ser asignados a un grupo determinado, sea el de estudio o el de control. Se diferencia de la selección al azar en que los individuos que se asignan pueden ser o no representativos de la población (sinónimo: asignación aleatoria).

**Asociación (Association).** Relación entre dos o más características u otras medidas, que es más intensa de lo que se esperaría solamente por azar. Cuando se usa para establecer el primer criterio de causa contribuyente, la asociación implica que las dos características aparecen en el mismo individuo con más frecuencia de la esperada exclusivamente por azar.

**Asociación de grupo (Group association).** Situación en la que una característica se presenta más frecuentemente en un grupo de individuos que en otro.

**Centro de gravedad de la distribución de las observaciones.** Forma especial de promedio.

**Coeficiente de correlación (Correlation coefficient).** Estadístico utilizado para estudiar la fuerza de una asociación entre dos variables, cada una de las cuales se ha extraído por muestreo de la población de interés mediante un método representativo o aleatorio.

**Coeficiente de determinación o cuadrado del coeficiente de correlación (Coefficient of determination, $R^2$)**. Estadístico que indica la proporción de la variabilidad de una variable (la variable dependiente), que es explicada conociendo un valor de una o más variables (las variables independientes).

**Cohorte (Cohort).** Grupo de individuos que comparten una exposición, una experiencia o una característica (véase estudio de cohorte, efecto de cohorte).

**Datos continuos (Continuous data).** Tipo de datos con un número ilimitado de valores espaciados uniformemente (por ejemplo, la tensión arterial diastólica, la colesterolemia).

**Datos nominales (Nominal data).** Aquellos datos que se dividen en categorías. Si los datos nominales, contienen más de dos categorías, estas no se pueden ordenar (por ejemplo, la raza o el color de los ojos). Los datos nominales necesitan más de una variable nominal si existen más de dos posibles categorías.

**Datos ordinales (Ordinal data).** Datos sobre un número limitado de categorías que tienen un orden inherente de menor a mayor. Sin embargo, los datos ordinales no predeterminan el espacio que existe entre las categorías (por ejemplo, estadíos 1, 2, 3 y 4 de un cáncer).

**Desviación estándar (Standar deviation).** Medida de la dispersión de los datos empleada habitualmente. El cuadrado de la desviación estándar se denomina «varianza» (sinónimo: desviación típica).

**Distribución (Distribution).** Frecuencias absolutas o relativas de todos los posibles valores de una característica. Las poblaciones y las distribuciones se pueden describir matemática o gráficamente. Uno de los objetivos de la estadística es estimar parámetros de las distribuciones poblacionales.

**Distribución gausiana (Gaussian distribution).** Una distribución de los datos que se supone en numerosas pruebas estadísticas. La distribución gausiana se representa en una curva simétrica, continua y acampanada, en la cual el valor de la media corresponde al punto más alto (sinónimo: distribución normal).

**Distribución normal (Normal distribution).** Véase distribución gausiana.

**Efecto (Effect).** Un desenlace o resultado que es producido, al menos en parte, por un factor etiológico conocido como «causa».

**Eficacia (Efficacy).** Grado en que un tratamiento produce un efecto beneficioso cuando se valora bajo las condiciones ideales de una investigación. La eficacia es al tratamiento lo que la causa contribuyente es a la etiología de la enfermedad.

**Error de muestreo (Sampling error).** Error introducido por las diferencias debido al azar entre la estimación obtenida en la muestra y el valor verdadero en la población de la que se ha extraído dicha muestra. El error de muestreo es inherente al uso de métodos de muestreo, y el error estándar cuantifica su magnitud.

**Error estándar (Standar error).** Grado de dispersión de las estimaciones puntuales obtenidas en muestras de un tamaño determinado.

**Estadístico (Statistic).** Valor calculado a partir de los datos de una muestra. Es utilizado para estimar un valor o parámetro de la población extraída de dicha muestra (sinónimos: valor muestral, estadígrafo).

**Estandarización de una tasa (Standardization, of a rate).** Proceso que permite tomar en cuenta o ajustar los datos según los efectos de un factor, como edad o sexo, sobre las tasas calculadas.

**Estimación (Estimate).** Valor o intervalo de valores calculados a partir de una muestra de observaciones, que se emplea como aproximación al valor correspondiente en la población; es decir, al parámetro.

**Estimación por intervalo (Interval estimate).** Véase intervalo de confianza.

**Estimación puntual (Point estimate).** Valor único calculado a partir de las observaciones muestrales, que se utiliza como estimación del valor poblacional o parámetro.

**Estratificación (Stratification).** División en grupos. El mismo término se puede emplear para hacer referencia al proceso de control, según las diferencias entre los factores de confusión, mediante la obtención de estimaciones separadas para los grupos de individuos que tienen los mismos valores de la variable de confusión. La estratificación también puede referirse a un método de muestreo intencionado y diseñado para representar en exceso categorías poco frecuentes de una variable independiente.

**Exactitud (Accuracy).** Capacidad de una prueba para producir resultados que se aproximen al verdadero valor del fenómeno. Falta de error sistemático o aleatorio; precisión sin sesgo.

**Experimento natural (Natural experiment).** Investigación en la cual la modificación de un factor de riesgo se produce en un grupo de individuos, pero no en un grupo de control. A diferencia del ensayo clínico controlado, en el experimento natural la modificación no es producida por la intervención del investigador.

**Extrapolación (Extrapolation).** Conclusiones sobre el significado del estudio para una población-objetivo compuesta por individuos o datos no representados en la muestra estudiada.

**Factor de riesgo (Risk factor).** Característica o factor asociado a un aumento de la probabilidad que aparezca una enfermedad. Un factor de riesgo no implica necesariamente la existencia de una relación causa-efecto. En este libro, el factor de riesgo implica que, al menos, se haya establecido una asociación a nivel individual.

**Grados de libertad (Degrees of freedom).** Parámetro de muchas distribuciones estadísticas estándares. Los grados de libertad permiten tomar en cuenta el número de parámetros poblacionales que se deben estimar en una muestra para poder aplicar ciertas pruebas estadísticas.

**Grupo de control (Control group).** Grupo de personas que se selecciona para comparación con el grupo de estudio. Idealmente, el grupo de control es idéntico al de estudio, excepto en que no posee la característica estudiada o no ha sido expuesto al tratamiento que se investiga (sinónimo: grupo de referencia o grupo testigo).

**Grupo de estudio (Study group).** En un estudio de cohortes o en un ensayo clínico controlado, este es el grupo de individuos que posee las características o está expuesto a los factores estudiados. En los estudios de casos y controles o en los transversales, corresponde al grupo de individuos que padecen la enfermedad investigada.

**Grupo de referencia (Reference group).** Grupo de individuos, presuntamente sanos, del que se extrae una muestra de sujetos en los que se realizarán mediciones para establecer un intervalo de normalidad (sinónimo: población de referencia).

**Hipótesis de estudio (Study hypothesis).** Afirmación de la existencia de una asociación entre dos o más variables en la población de la que procede la muestra. Una hipótesis de estudio es unilateral cuando solo considera las asociaciones en una dirección; es bilateral cuando no se especifica la dirección de la asociación.

**Hipótesis nula (Null hypothesis).** Afirmación de que no existe una asociación o diferencia verdadera entre las variables en la población de la que se extrajo la muestra estudiada.

**Inferencia (Inference).** En términos estadísticos, la inferencia es el proceso lógico que tiene lugar durante las pruebas de significación estadística (véase prueba de significación estadística).

**Interpretación (Interpretation).** Extracción de conclusiones sobre el significado de cualquier diferencia observada entre el grupo de estudio y el de control incluidos en la investigación.

**Intervalo de confianza de 95% (Confidence interval 95%).** En términos estadísticos, es el intervalo de valores numéricos donde se encuentra el valor poblacional que se está estimando, con un nivel de confianza de 95% (sinónimo: estimación por intervalo).

**Intervalo de normalidad (Range of normal).** Medida del intervalo de valores obtenidos en una prueba, correspondientes a los sujetos que padecen la enfermedad. Con frecuencia, hace referencia al 95% de los valores centrales o a la media de los valores de los individuos sin la enfermedad, más o menos dos desviaciones estándares (sinónimo: valores normales 0).

**Media (Mean).** Suma de todas las mediciones, dividida por el número total de valores sumados.

**Mediana (Median).** Punto medio de la distribución. La mediana es el valor que deja la mitad de los valores por arriba y la otra mitad por debajo.

**Muestra (Sample).** Subgrupo de una población obtenido por un investigador para extraer conclusiones o para realizar estimaciones sobre la población.

**Muestra aleatoria (Naturalistic sample).** Grupo de observaciones obtenidas de una población, de forma tal que la distribución muestral de los valores de la variable independiente es representativa de su distribución en la población (sinónimo: muestra al azar).

**Muestra fortuita (Chunk sample).** Muestra que se extrae de una población representativa de dicha población, por lo fácil que resulta obtener datos de ella sin tener en cuenta el grado (aleatoria o representativa) de dicha población.

**Muestra intencionada (Purposive sample).** Grupo de observaciones obtenidas a partir de una población; de forma tal que el investigador determina la distribución muestral de los valores de la variable independiente, sin que sea necesariamente representativa de su distribución en la población.

**No sesgado (Unbiased).** Sin error sistemático asociado.

**Parámetro (Parameter).** Valor que sintetiza la distribución de una población. Uno de los objetos del análisis estadístico consiste en estimar los parámetros poblacionales a partir de las observaciones muestrales (sinónimo: valor poblacional).

**Población (Population).** Grupo numeroso compuesto con frecuencia, pero no necesariamente por individuos. En estadística, el objetivo es extraer conclusiones acerca de una o más poblaciones mediante la obtención de subgrupos o muestras compuestas por individuos pertenecientes a la población.

**Población-objetivo (Target population).** Grupo de individuos a los que se desea extrapolar o aplicar los resultados de una investigación. La población objetivo puede ser (y de hecho lo es) frecuentemente distinta de la población de la que se extrae la muestra de una investigación.

**Potencia (Power).** Capacidad de un estudio para demostrar significación estadística cuando existe una diferencia o una asociación verdadera de una fuerza determinada, en la población de la que se ha extraído la muestra (sinónimo: poder estadístico, poder de resolución).

**Preciso (Precise).** Sin error aleatorio asociado (una medición imprecisa puede desviarse del valor numérico verdadero en cualquier dirección).

**Probabilidad (Probability).** Proporción en la cual el numerador es el número de veces que ocurre un suceso, y el denominador ese mismo número sumado al número de veces que no ocurre ese suceso.

**Proporción (Proportion).** Fracción cuyo numerador está formado por un subgrupo de individuos incluidos en el denominador.

**Prueba bilateral (Two-tailed test).** Prueba de significación estadística en la que se toman en cuenta las desviaciones de la hipótesis nula en cualquier dirección. El uso de una prueba bilateral implica que el investigador deseaba considerar las desviaciones en cualquier dirección antes de recoger los datos (sinónimo: prueba de dos colas).

**Prueba de significación estadística (Statistical significance test).** Técnica estadística para calcular la probabilidad de que la asociación observada en una muestra hubiera podido ocurrir por azar, si no existiese esa asociación en la población origen (sinónimo: inferencia, contraste de hipótesis).

**Prueba unilateral (One tailde test).** Prueba de significación estadística en la cual solo se toman en cuenta las desviaciones, respecto de la hipótesis nula en una sola dirección. El empleo de una prueba unilateral implica que el investigador no considera posible una desviación verdadera en dirección opuesta (sinónimo: prueba de una cola).

**Pruebas estadísticas no paramétricas (Nonparametric statistics).** Pruebas estadísticas en las que no existen supuestos sobre la distribución de los parámetros en la población de la que se extrajo la muestra. En estas pruebas se aceptan otros supuestos, como el relativo a la aleatoriedad del muestreo. Se aplican con mayor frecuencia a los datos ordinales, si bien pueden emplearse también para analizar datos continuos transformados en una escala ordinal (sinónimo: pruebas).

**Razón (Ratio).** Fracción en la cual el numerador no es necesariamente un subconjunto del denominador, como ocurre en la proporción.

**Razón de mortalidad estandarizada (Standardized mortality ratio).** Fracción cuyo numerador es el número de muertes observadas y cuyo denominador corresponde al número de muertes esperadas sobre la base de una población de referencia. La razón de mortalidad estandarizada implica que, para ajustar los datos según los factores de confusión, se ha empleado la estandarización indirecta.

**Razón de mortalidad proporcional (Proportionate mortality ratio).** Fracción cuyo numerador está formado por el número de personas que mueren de una enfermedad concreta durante un periodo determinado, cuyo denominador es el número de individuos fallecidos por todas las causas en el mismo periodo.

**Recorrido (Range).** Diferencia entre los valores máximo y mínimo de una población o de una muestra (sinónimo: amplitud).

**Regresión a la media (Regression to the mean).** Principio estadístico que indica que es improbable que los sucesos infrecuentes vuelvan a suceder. Es más probable, solo por azar, que las mediciones siguientes a un resultado infrecuente sean más cercanas a la media. Además, es posible que existan factores psicológicos o sociales que contribuyan a forzar los sucesos posteriores a "regresar" hacia valores más cercanos a la media.

**Reproducible (Reproducibility).** Capacidad de una prueba para producir resultados consistentes cuando se repite en condiciones similares y se interpreta sin conocimiento de los resultados previos (sinónimo: fiabilidad, repetibilidad).

**Riesgo (Risk).** Probabilidad de que ocurra un suceso durante un periodo determinado. El numerador del riesgo es el número de individuos donde aparece la enfermedad durante dicho periodo, mientras que el denominador es el número de sujetos sin la enfermedad al inicio del periodo.

**Riesgo relativo (Relative risk).** Razón entre la probabilidad de que suceda un desenlace en un periodo determinado en los expuestos al factor de riesgo, y la probabilidad de que suceda entre los no expuestos al factor de riesgo en el mismo periodo. El riesgo relativo es una medida de la fuerza o del grado de asociación aplicable a los estudios de cohorte y a los ensayos clínicos aleatorios. En los casos y controles, la razón de productos cruzados se puede utilizar frecuentemente como una aproximación al riesgo relativo.

**Robusto (Robust).** Se dice que una prueba estadística es robusta si se puede violar sus supuestos sin que ello repercuta sustancialmente en las conclusiones.

**Selección al azar (Random selection).** Método para obtener una muestra que asegura que cada individuo de la población tiene una probabilidad conocida, aunque no necesariamente igual, de ser seleccionado para formar parte de la muestra.

**Sesgo (Bias).** Un factor que produce la desviación sistemática de un resultado en una dirección, en relación con los valores reales. El uso de este término está limitado a las desviaciones originadas por defectos en el diseño del estudio.

**Tasa (Rate).** Habitualmente, se emplea para indicar cualquier medida de la frecuencia de la enfermedad o desenlace. Desde un punto de vista estadístico, las tasas son aquellas medidas de la frecuencia de la enfermedad que incluyen una medida de tiempo en el denominador (por ejemplo, la incidencia).

**Tasa de mortalidad (Mortality rate).** Es una medida de la incidencia de muerte. Esta tasa se calcula dividiendo el número de muertes que han ocurrido durante un periodo, por el producto del número de individuos y el número de unidades de tiempo del periodo de seguimiento.

**Técnicas de regresión (Regression techniques).** Métodos estadísticos útiles para describir la asociación entre una variable dependiente y una o más variables independientes. Las técnicas de regresión se emplean con frecuencia para ajustar el efecto, según las variables de confusión.

**Válido (Valid).** Una medición es válida si es apropiada para la cuestión que se está investigando, o si mide lo que intenta medir.

**Valor P (P value).** Probabilidad de realizar una observación, tan alejada de la condición descrita en la hipótesis nula como la observada en su conjunto de datos, si la hipótesis nula fuera cierta. El cálculo del valor P constituye el "objetivo" de las pruebas de significación estadística.

**Variabilidad interobservador (Interobserver variability).** Variabilidad en las medidas realizadas por diversos observadores.

**Variabilidad intraobservador (Intraobserver variability).** Variabilidad en las medidas realizadas por el mismo observador en distintas ocasiones.

**Variable (Variable).** En su acepción general, variable se refiere a una característica que se mide en el estudio. En términos estadísticos rigurosos, una variable es la representante de esas mediciones en el análisis. Los datos medidos en una escala continua u ordinal se expresan por medio de una variable, como ocurre con las variables nominales que solo tienen dos categorías. Sin embargo, los datos nominales con más de dos categorías deben expresarse con más de una variable.

**Variable de confusión (Confounding variable).** Característica o variable que se distribuye de forma diferente en el grupo de estudio y en el de control; además, afecta al desenlace estudiado. Una variable de confusión puede deberse al azar o a un sesgo. Cuando se debe a un sesgo en el proceso de asignación, el error resultante se denomina «sesgo de selección» (sinónimo: factor de confusión).

**Variable dependiente (Dependent variable).** En general, la variable del desenlace de interés en cualquier tipo de estudio. El desenlace o resultado que uno pretende explicar o estimar.

**Variable independiente (Independent variable).** Variable que se mide para determinar el valor correspondiente de la variable dependiente en cualquier tipo de estudio. Las variables independientes definen las condiciones bajo las cuales se examinará la variable dependiente.

**Varianza (Variance).** Véase desviación estándar.

## APÉNDICE

Notación científica:
En ocasiones, en las tablas de resultados del SPSS se incluyen datos en la denominada «notación científica», debido a que los números escritos de esta forma ocupan menos espacio que en la notación tradicional. Por ejemplo:

**Estadísticos descriptivos**

| | N | Media | Varianza | Desv. tip. |
|---|---|---|---|---|
| ESTATURA | 200 | 1.6704 | 6.235E-03 | 7.896E-02 |
| PESO | 200 | 57.89 | 92.752 | 9.63 |

La notación científica se usa para señalar la posición del primer valor decimal. Por ejemplo, la varianza para la variable *Estatura* aparece escrita en notación científica: 6.235E-03. Esta expresión significa lo siguiente en notación tradicional:

$$6.235E-03 = 6.235 \times 10^{-3} = \frac{6.235}{1000} = 0.006325$$

Del mismo modo, el número 5.274E+04 es igual a 5274000.

## BIBLIOGRAFÍA


Camacho, J. (2000). *Estadística con SPSS para Windows*. Madrid: RA-MA.

Corston, R. y Colman, A. (2000). *A Crash Course in SPSS for Windows*. Oxford: Blackwell Publishers.

Manzano, V.; Varela, J.; García, A. y Pérez F. J. (1999). *SPSS para Windows*. Madrid: Ed. Rama.

Miquel, S.; Bigné, E.; Cuenca, A.; Miquel, J. y Lévy, J. (1999). *Investigación de mercados*. Madrid: McGraw Hill.

Palmer, A.L. (1999). *Análisis de datos. Etapa exploratoria, psicología*. Madrid: Ed. Pirámide.

Pérez, C. (2001). *Técnicas estadísticas con SPSS*. Madrid: Prentice Hall.

Peña D. (2002). *Análisis de datos multivariantes*. McGraw-Hill Interamericana de España.

Quezada, N. (2010). *Estadística para ingenieros*. Lima: Empresa Editora Macro.

Quezada, N. (2010). *Metodología de la investigación*. Lima: Empresa Editora Macro.

SPSS Inc. (2000). *SPSS 10.0. Manual de usuario, SPSS*. Chicago.


## Fuentes electrónicaas consultadas:


http://www.spss.com/
http://www.spss.com/es/
http://www.amstat.org/index.cfm?fuseaction=main
http://www.amstat.org/publications/jasa/index.cfm?fuseaction=main
http://www.aulafacil.org/CursoEstadistica/CursoEstadistica.htm
http://www.estadistico.com/

Impreso en los Talleres Gráficos de




Surquillo