ESTADÍSTICA PARA LA TESIS DE POSTGRADO

Jully Pahola Calderón Saldaña Ph.D.
Luis Alex Alzamora de los Godos Urcia Ph. D.

SPSS APLICADO A LA TESIS DE POSTGRADO

Una forma simple de hacer estadística con el Programa SPSS

JULLY PAHOLA CALDERÓN SALDAÑA Ph. D.
LUIS ALZAMORA DE LOS GODOS URCIA Ph. D.

Lima, 2011

ISBN #: 978-1-257-86896-4

Editorial: LULU International
Registrado en Inglaterra: Nº 5720154

LULU
3101 HILLSBOROUGH STREET
RALEIGH, NC 27607
North Carolina
UNITED STATES
www.lulu.com\content/9882508

ÍNDICE

INTRODUCCIÓN

El Libro de SPSS aplicado a la tesis de postgradi cambia el paradigma de la estadística teórica y se concentra en convertirla en un sistema de aprendizaje práctico y útil, ya que el candidato a maestro no debe ser un matemático estadístico, sino un profesional de la educación que emplee la estadística de manera práctica y sencilla para trabajar su tesis, en este sentido se le facilitará los elementos teórico prácticos para utilizar un programa informático para llevar a cabo el tratamiento y análisis de información estadística. Se dirige a un conjunto muy amplio de estudiantes de maestría, tanto aquellos que se inicien en el aprendizaje de la Estadística como para los que ya tienen unos conocimientos previos sobre la materia y quieren aplicarlos con la ayuda de un programa ampliamente difundido en la actualidad como es el programa SPSS.

Se presupone que el participante de maestría que utiliza esta aplicación quiere introducirse en los conocimientos de la Estadística mediante la utilización de un programa informático para el tratamiento de datos, concretamente el programa SPSS, versión 11. Para el seguimiento del módulo no se requiere ningún conocimiento previo del funcionamiento de este programa, y muy pocos conocimientos de matemática básica. Este material ha sido concebido como un instrumento aplicado al aprendizaje de la Estadística, ya que permite ver cómo se aplican los conocimientos y se obtienen los resultados con las herramientas informáticas disponibles.

En cada uno de los apartados se consideran dos partes que permiten, en primer lugar, familiarizarse con el entorno del programa SPSS, y seguidamente se procede a explicar las técnicas de análisis de datos: se incluyen una explicación teórica con definiciones, expresiones y fórmulas que permite introducir o recordar al lector la teoría estadística que se está utilizando.

Al finalizar el trabajo de este material, el usuario habrá adquirido los conocimientos necesarios para utilizar el programa SPSS en los siguientes aspectos:
- Introducción y lectura de los datos.

- Análisis de estadística descriptiva básica univariante.
- Tablas de frecuencias bivariantes.
- Contraste de hipótesis paramétricas y no paramétricas.
- Especificación, estimación y evaluación de un modelo de regresión lineal simple.
- Identificación de modelos de series temporales y realización de predicciones.

Este material tiene un enfoque eminentemente práctico dado que para cada uno de los procesos incluidos se presentan: instrucciones de los pasos a seguir, imágenes de las pantallas que se van obteniendo y ejemplos resueltos incluyendo los resultados obtenidos por el programa, así como todas las fases intermedias que nos llevan a ellos, y las conclusiones que pueden extraerse de los mismos.

PROGRAMACIÓN ACADÉMICA

OBJETIVOS

OBJETIVOS DEL LIBRO

- Emplear adecuadamente la decisión de las pruebas estadísticas.
- Aplicar instrumentos, construir una base de datos y realizar el análisis estadístico pertinente.

OBJETIVOS ESPECÍFICOS

- Identificar las pruebas estadísticas pertinentes al tema de tesis.
- Procesar los instrumentos de recolección de datos.
- Construir una base de datos.
- Realizar la estadística descriptiva.
- Desarrollar análisis de datos para variables cualitativas y cuantitativas.

Primera unidad

.....

Construcción la base de datos

1.1 INTRODUCCIÓN AL SPSS

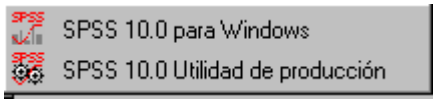
Objetivo: El alumno al terminar el aprendizaje del presente capítulo tendrá la capacidad de conocer el programa SPSS y manejar el editor de datos.

INTRODUCCIÓN

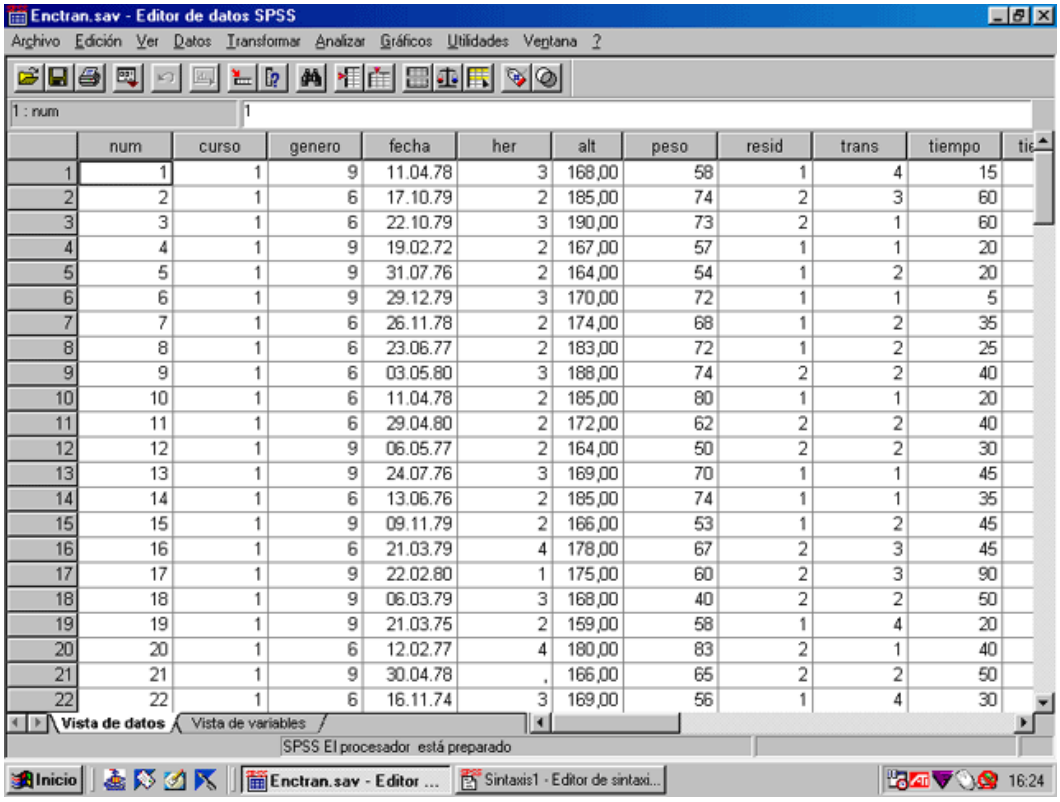
El SPSS es un programa de análisis estadístico fácil de utilizar y con gran capacidad operativa. Permite analizar datos almacenados en diversos formatos y generar documentos con alta calidad de presentación.

EDITOR DE DATOS

Al instalar el programa se crean, automáticamente, los siguientes iconos que aparecen en la barra Programas de Windows.



El icono *SPSS para Windows* da acceso al programa. Seleccionándolo con el cursor se entra en el programa y aparece la ventana *Editor de datos*.



	num	curso	genero	fecha	her	alt	peso	resid	trans	tiempo
1	1	1	9	11.04.78	3	168,00	58	1	4	15
2	2	1	6	17.10.79	2	185,00	74	2	3	60
3	3	1	6	22.10.79	3	190,00	73	2	1	60
4	4	1	9	19.02.72	2	167,00	57	1	1	20
5	5	1	9	31.07.76	2	164,00	54	1	2	20
6	6	1	9	29.12.79	3	170,00	72	1	1	5
7	7	1	6	26.11.78	2	174,00	68	1	2	35
8	8	1	6	23.06.77	2	183,00	72	1	2	25
9	9	1	6	03.05.80	3	188,00	74	2	2	40
10	10	1	6	11.04.78	2	185,00	80	1	1	20
11	11	1	6	29.04.80	2	172,00	62	2	2	40
12	12	1	9	06.05.77	2	164,00	50	2	2	30
13	13	1	9	24.07.76	3	169,00	70	1	1	45
14	14	1	6	13.06.76	2	185,00	74	1	1	35
15	15	1	9	09.11.79	2	166,00	53	1	2	45
16	16	1	6	21.03.79	4	178,00	67	2	3	45
17	17	1	9	22.02.80	1	175,00	60	2	3	90
18	18	1	9	06.03.79	3	168,00	40	2	2	50
19	19	1	9	21.03.75	2	159,00	58	1	4	20
20	20	1	6	12.02.77	4	180,00	83	2	1	40
21	21	1	9	30.04.78	.	166,00	65	2	2	50
22	22	1	6	16.11.74	3	169,00	56	1	4	30

La ventana *Editor de datos* permite gestionar la entrada, lectura, transformación, importación y almacenaje de ficheros de datos.

El editor está formado por un conjunto de filas y columnas en las que se visualizan los datos del archivo activo. Las columnas recogen las variables del archivo, las filas los individuos o elementos observados y las celdas los valores.

Además el editor presenta las siguientes barras:

1. *Barra de menú* del editor.

• *Archivo*: presenta los procedimientos relacionados con la lectura, impresión y almacenaje de archivos.

• *Edición*: contiene las opciones de copiar, mover y pegar del entorno Windows.

• *Ver*: modifica la visualización de las barras y pantalla.

• *Datos*: permite definir variables y modificarlas bien temporalmente o bien de manera definitiva; en este caso se deberá salvar el archivo antes de finalizar la sesión.

• *Tranformar*: permite definir, temporalmente o de manera definitiva, nuevas variables a partir de las existentes.

• *Analizar*: recoge los procedimientos estadísticos.

• *Gráficos*: permite la creación, modificación y edición de una amplia gama de gráficos.

• *Utilidades*: informa sobre las características de los archivos de datos.

• *Ventana*: presenta las opciones de ventana del entorno Windows.

• *?*: Permite consultar la ayuda o el tutorial.

2. *Barra de herramientas*. Contiene un conjunto de iconos que dan acceso directo a algunos procedimientos.

• Los tres primeros iconos activan las opciones abrir, guardar e imprimir, respectivamente, del menú Archivo y permiten, tal como indican sus nombres, abrir, almacenar e imprimir el archivo de datos.

•  Da acceso a los últimos cuadros de diálogo utilizados.

•  Deshace la última modificación.

•  Permite ir a un gráfico determinado.

•  Desplazan el cursor a la fila (n\circ de individuo o elemento muestral) o a la columna (variable) indicada, respectivamente.

•  Busca, en la variable seleccionada, un dato.

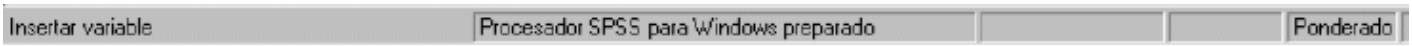
•  Añaden una fila (elemento) o una columna (variable), respectivamente.

•  El primero segmenta el archivo, el segundo permite activar un criterio de ponderación y el tercero selecciona casos a analizar.

•  Muestra u oculta las etiquetas de los valores de las variables.

•  Permite usar conjuntos de variables previamente definidos.

3. *Barra de estado*. Se encuentra en la parte inferior de la pantalla e indica el estado actual del proceso, el número de elementos que se están procesando, las iteraciones realizadas y los filtrados, ponderaciones o segmentaciones activados.



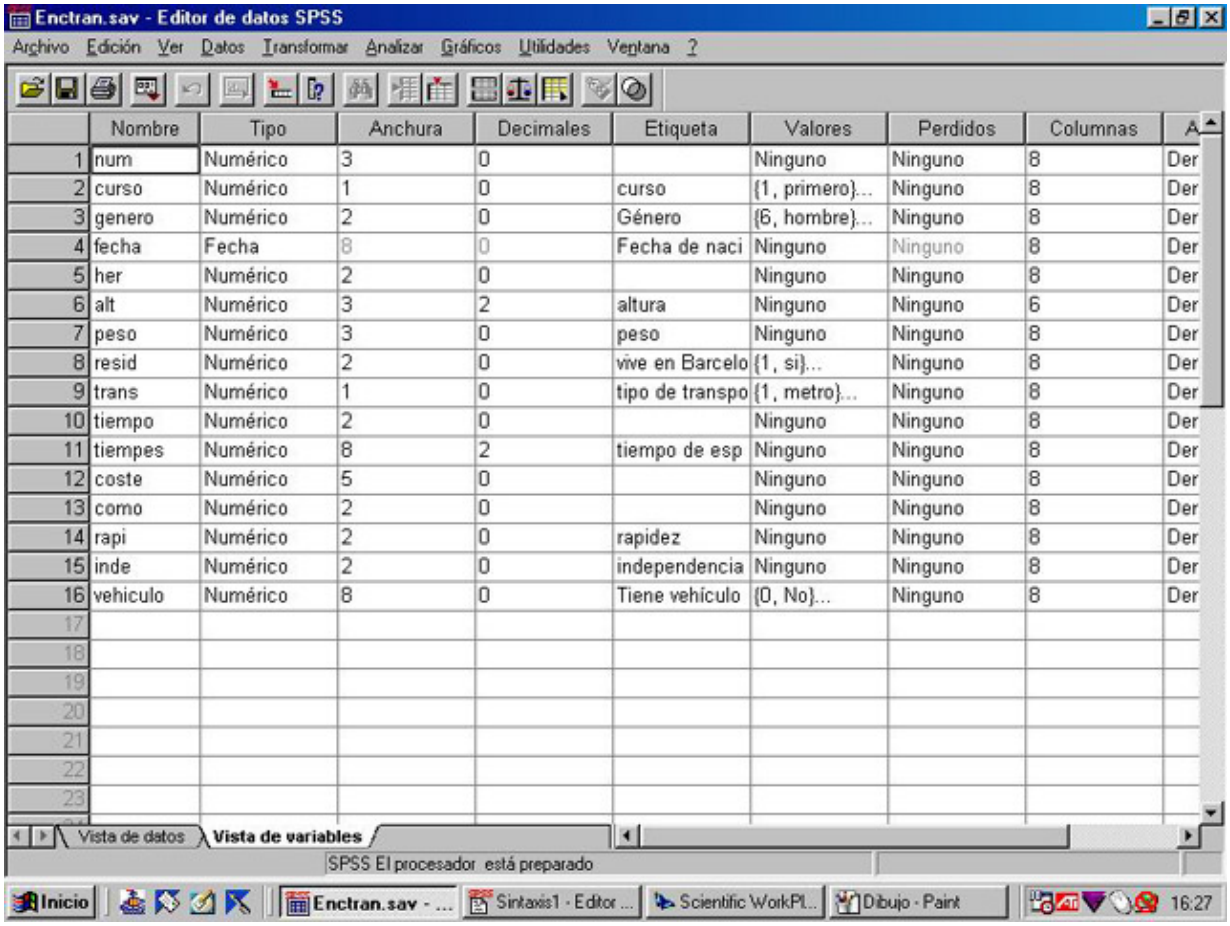
CREAR UN ARCHIVO DE DATOS

Cuando se inicia la sesión con SPSS aparece el *editor de datos* en blanco, ya reparado para crear las variables y entrar sus correspondientes valores.

1. *Definir las variables*. Antes de introducir los datos es preciso definir las variables, es decir, especificar el nombre de la variable y el tipo de datos que contendrá cada columna.

Puede hacerse con opciones de menú: *Datos e Insertar variable*

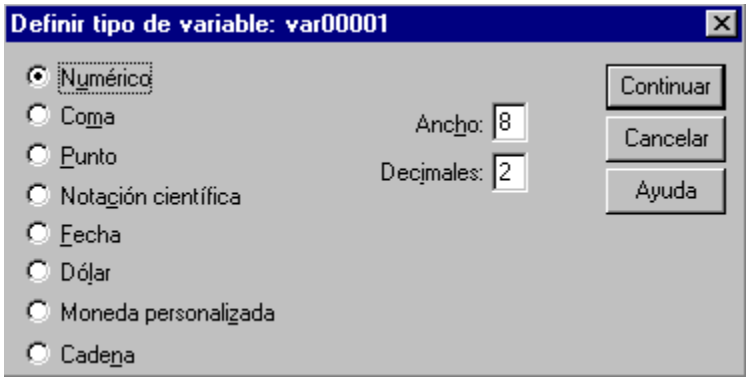
con lo que se define la variable con las características por defecto o activando *Vista de variables*.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	A
1	num	Numérico	3	0		Ninguno	Ninguno	8	Der
2	curso	Numérico	1	0	curso	{1, primero}...	Ninguno	8	Der
3	genero	Numérico	2	0	Género	{6, hombre}...	Ninguno	8	Der
4	fecha	Fecha	8	0	Fecha de naci	Ninguno	Ninguno	8	Der
5	her	Numérico	2	0		Ninguno	Ninguno	8	Der
6	alt	Numérico	3	2	altura	Ninguno	Ninguno	6	Der
7	peso	Numérico	3	0	peso	Ninguno	Ninguno	8	Der
8	resid	Numérico	2	0	vive en Barcelo	{1, si}...	Ninguno	8	Der
9	trans	Numérico	1	0	tipo de transpo	{1, metro}...	Ninguno	8	Der
10	tiempo	Numérico	2	0		Ninguno	Ninguno	8	Der
11	tiempes	Numérico	8	2	tiempo de esp	Ninguno	Ninguno	8	Der
12	coste	Numérico	5	0		Ninguno	Ninguno	8	Der
13	como	Numérico	2	0		Ninguno	Ninguno	8	Der
14	rapi	Numérico	2	0	rapidez	Ninguno	Ninguno	8	Der
15	inde	Numérico	2	0	independencia	Ninguno	Ninguno	8	Der
16	vehiculo	Numérico	8	0	Tiene vehículo	{0, No}...	Ninguno	8	Der
17									
18									
19									
20									
21									
22									
23									

El *Nombre* de la variable debe empezar por una letra y como máximo puede tener 8 caracteres. No puede contener espacios en blanco, caracteres especiales (¡ , ?, ', *) ni palabras clave SPSS (ALL, AND, NOT, OR...). Lógicamente no puede asignarse el mismo nombre a dos o más variables.

El Tipo de variable por defecto es numérico con 8 dígitos, 2 de ellos decimales. Para modificarlo se debe seleccionar el botón lateral y se accede al cuadro de diálogo *Tipo de variable*:4



Los tipos más frecuentes son:

• *Numérico*: presenta los decimales separados por un punto o coma dependiendo de la configuración numérica del ordenador.

• *Coma*: presenta los millares separados por una coma y los decimales por un punto.

• *Punto*: presenta los millares separados por un punto y los decimales por una coma.

• *Fecha*: abre un amplio directorio de formatos fecha.

• *Cadena*: recoge variables no numéricas y alinea a la izquierda sus categorías.

La columna *Etiqueta* permite asignar una descripción a la variable. Las etiquetas de las variables no pueden exceder de 120 caracteres.

La columna *Valores* permite asignar etiquetas a los valores de la variable. Éstos son especialmente útiles cuando la variable es categórica y se emplean números para representar las categorías. Por ejemplo: 0 'hombre' 1 'mujer'.

La columna *Valores perdidos* abre un cuadro de diálogo que permite especificar el código de valores missing que se asigna a la variable.

En *Columna* y *Alineación* se puede modificar el ancho y la alineación del contenido de la columna.

Por último, se puede especificar la escala de medida de la variable eligiendo entre: *Escala*, *Ordinal* y *Nominal*.

2. *Entrar los datos*: Los datos pueden ser tecleados en el orden que más convenga: por filas (elementos) o por columnas (variables). Para desplazar el cursor a la fila siguiente se debe pulsar la tecla Enter, mientras que para desplazarlo a la siguiente columna la tecla es *Tab*. Si se quiere acceder de forma rápida a una fila o columna determinada se pueden utilizar los iconos:

ALMACENAR EL ARCHIVO

Para almacenar un archivo creado o modificado se seleccionará el icono *Guardar* o bien las opciones del menú del editor de datos:

Archivo

Guardar

Ambos procedimientos abren el cuadro de diálogo *Guardar* donde se debe indicar el nombre del archivo, la carpeta y unidad donde se quiere almacenar y, opcionalmente, se puede seleccionar el tipo de archivo que se quiere guardar. Por defecto se almacena el archivo en formato SPSS (*.sav). Este formato sólo puede ser leído por versiones posteriores a la 6.

Si el archivo estaba creado al iniciar la sesión, al activar *Guardar* se almacena automáticamente sin que aparezca el cuadro de diálogo.

Si se quiere renombrar, cambiar de formato o reubicar un archivo ya creado al iniciar la sesión se debe activar la opción del menú del editor:

Archivo

Guardar como

En *Nombre de archivo* se puede renombrar. En *Guardar en* se puede modificar la carpeta o directorio donde se almacenará. Si se quiere salvar con otro formato en *Guardar como tipo* se debe seleccionar el deseado entre: versiones de SPSS anteriores, ASCII separado por tabuladores, ASCII formato fijo, Excel, diversas versiones de Lotus y de dBase.

EJEMPLO

Si cuenta con un cuestionario, con esta información se desea crear una nueva base de datos SPSS.

Los pasos a seguir son:

1. Entrar en el programa. *Doble clic sobre el icono SPSS*. Aparece el Editor de datos en blanco donde se puede empezar a definir las variables.

2. Definir las variables: Número de encuesta, género, fecha de nacimiento, número de hermanos, altura y peso; con la secuencia *Datos* > *Insertar Variables* del menú del Editor.

En la pestaña *Vista de variable* modificamos las siguientes características:

1. Número de Encuesta.
    a. Nombre: Num
    b. Tipo: Numérico; Anchura: 3; Decimales: 0

c. Etiqueta: Número de Encuesta

• Género.

Nombre: Genero

Tipo: Cadena; Caracteres: 1

Etiqueta: Género; Valores: Valor: 0; Etiqueta de valor: Hombre; y Añadir. Valor:1;

Etiqueta de valor: Mujer.

• Fecha de nacimiento.

Nombre: Fecha

Tipo: Fecha y seleccionar: dd.mm.aa

Etiqueta: Fecha de nacimiento

• Número de hermanos.

Nombre: Her

Tipo: Numérico; Anchura: 2

Etiqueta: Número de hermanos.

• Altura.

Nombre: Alt

Tipo: Numérico; Anchura: 3

Etiquetas: Altura.

• Peso.

Nombre: Peso

Tipo: Numérico; Anchura: 2

3. Salvar el archivo. Pulsando el icono *Guardar* o la opción del menú *Archivo* > *Guardar*

• *Guardar en*: C:\ Mis documentos

• *Nombre de archivo*: Ejemplo 1

• *Guardar como tipo*: mantener el formato SPSS .

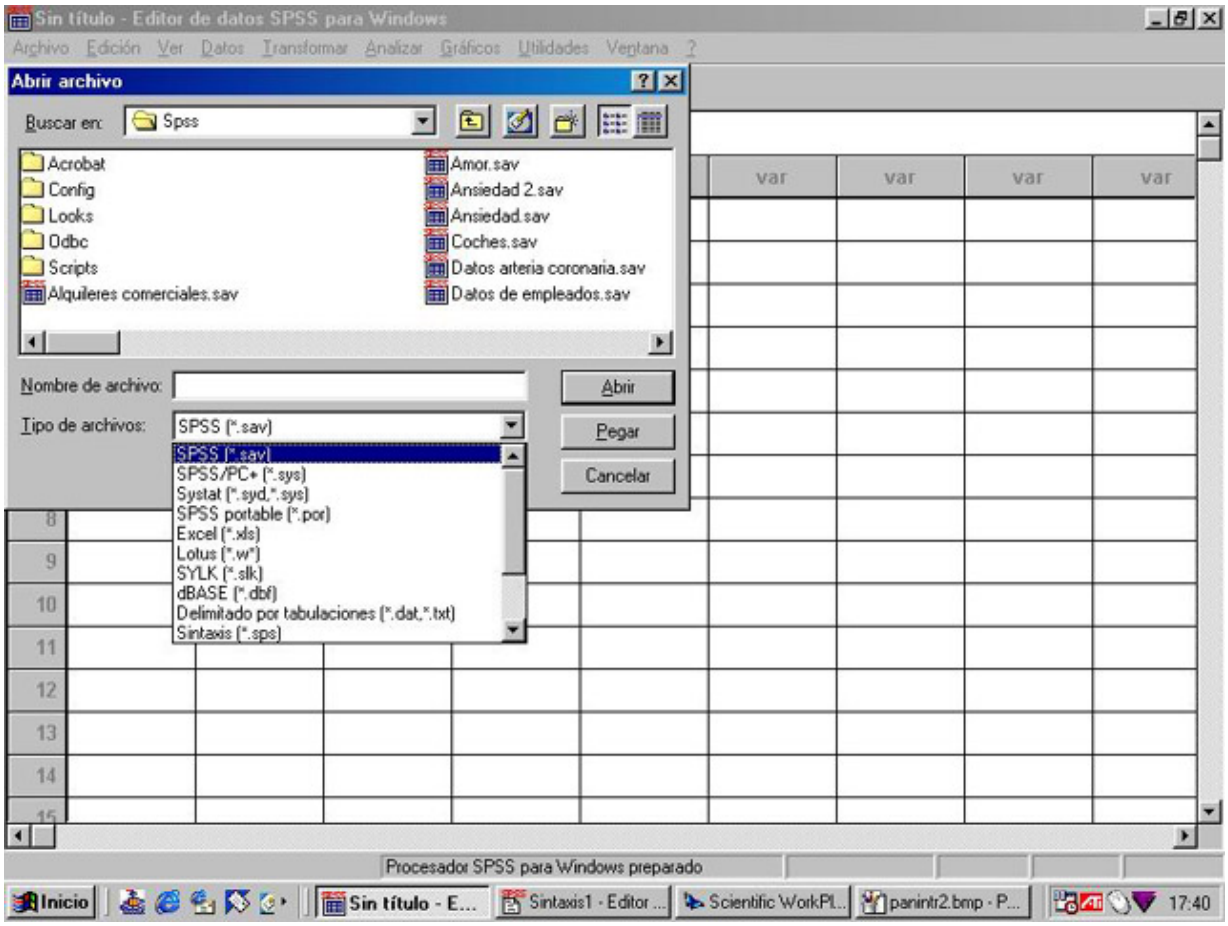
4. Introducir los valores. Situando el cursor en la primera celda de la primera fila se van introduciendo los valores por columnas pulsando Enter para cambiar de celda. Si se prefiere introducir los valores por filas, apretar Tab para cambiar de celda.

5. Guardar el archivo. Mediante el icono *Guardar* o con la opción del menú *Archivo* > *Guardar*

ABRIR UN ARCHIVO

Para abrir un archivo de datos SPSS (*.sav), SPSS/PC (*.sys), SYSTAT (*.sys), Excel (*.xls)*, Lotus (*.wk*), dBase (*.dbf) o un archivo ASCII con los valores separados por tabuladores, en el *Editor de datos* se activa la opción de menú:

Archivo
Abrir
Datos
o con el icono Abrir.



---

*Versiones 4.0 o anteriores

En pantalla, automáticamente, aparecen los archivos SPSS (*.sav) de la carpeta seleccionada. Se puede visualizar el contenido de otra carpeta u otra unidad seleccionándola en *Buscar en*. Para que liste el resto de archivos se deberá seleccionar su extensión en *Tipo de archivo*. El archivo se abrirá con doble clic o pulsando *Abrir*.

EJEMPLO

Se puede importar un archivo en el formato Excel. Se trata de leer este archivo y de realizar las siguientes modificaciones, pase en un disquete un archivo en Excel que

tenga una base de datos ordenadas, sin adornos, ni ediciones especiales, solamente una base ordenada.

1. *Abrir* el archivo. Icono *Abrir* o menú *Archivo* > *Abrir* > *Datos*:

• Buscar en: A:\

• Tipo de archivo: Excel

• Seleccionar con el ratón el archivo Enctran.xls y doble clic o Abrir.

• En el cuadro de diálogo Opciones de apertura de archivos seleccionar Leer nombres de variables.

En *Vista de variables*:

1. Etiquetar las siguientes variables.

Variable	*Etiqueta*
Num	Número de Encuesta
Her	Número de hermanos
Alt	Estatura
Como	Comodidad

2. Etiquetar los valores de las siguientes variables.

Variable	*Valor*	*Etiqueta*
Genero	6	Hombre
	9	Mujer
Resid	1	Si
	2	No

3. Modificar el tipo de variable.

Variable	*Tipo*	*Ancho*	*Decimales*
Curso	Numérico	2	0
Fecha	Fecha	dd-mmm-aaaa	
Coste	Punto	10	2

5. Establecer un ancho de columna en *Columnas* para la variable Curso igual a 2 y para la variable Coste igual a 10.

6. Modificar los siguientes valores en *Vista de datos*.

Variable	Buscar	Nuevo
Rapi	12	2
Inde	-1	3

Se sitúa el cursor sobre la columna de la variable que se quiere modificar (Rapi). Con la opción *Edición* > *Buscar* o el icono *Buscar* se desplaza el cursor hasta la celda correspondiente al valor (12) y se teclea el nuevo valor (2).

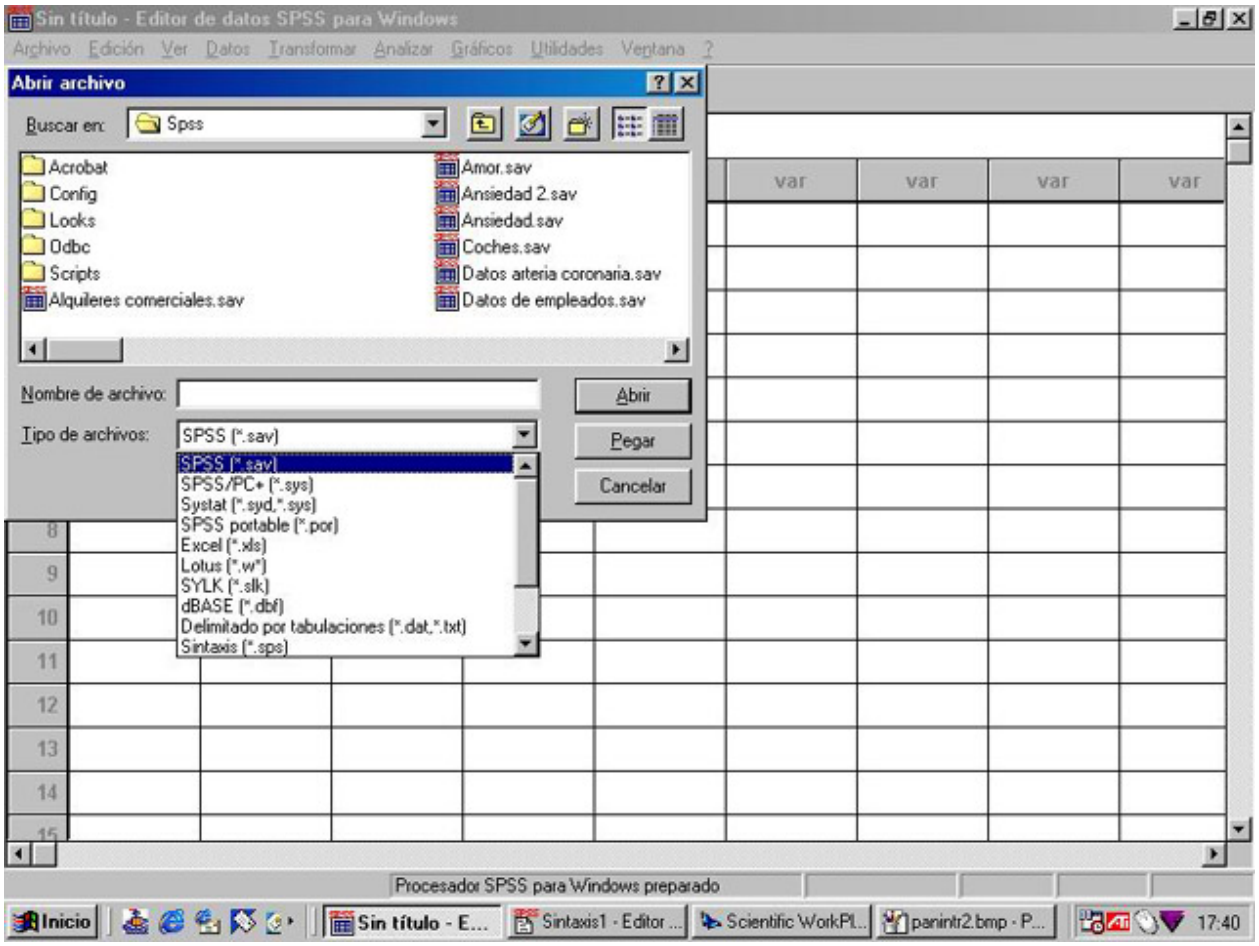
Se procede de la misma forma con la variable Inde. Se sitúa el cursor sobre la columna de esta variable. Con la opción *Edición >Buscar* o el icono Buscar se desplaza el cursor hasta la celda correspondiente al valor (-1) y se teclea el nuevo valor (3).

7. Guardar el archivo de datos con formato SPSS extensión SAV y nombre **Ejemplo**.

ABRIR UN ARCHIVO

Para abrir un archivo de datos SPSS (*.sav), SPSS/PC (*.sys), SYSTAT (*.sys), Excel (*.xls)*, Lotus (*.wk*), dBase (*.dbf) o un archivo ASCII con los valores separados por tabuladores, en el *Editor de datos* se activa la opción de menú:

Archivo
Abrir
Datos
o con el icono *Abrir*.



*Versiones 4.0 o anteriores

En pantalla, automáticamente, aparecen los archivos SPSS (*.sav) de la carpeta seleccionada. Se puede visualizar el contenido de otra carpeta u otra unidad

seleccionándola en *Buscar en*. Para que liste el resto de archivos se deberá seleccionar su extensión en *Tipo de archivo*.

El archivo se abrirá con doble clic o pulsando *Abrir*.

EJEMPLO

La información correspondiente a un archivo con formato Excel. Se trata de leer este archivo y de realizar modificaciones.

1. *Abrir* el archivo. Icono *Abrir* o menú *Archivo* > *Abrir* > *Datos*:

• *Buscar en*: A:\
• *Tipo de archivo*: Excel
• Seleccionar con el ratón el archivo **.xls** y doble clic o *Abrir*.
• En el cuadro de diálogo *Opciones de apertura* de archivos seleccionar *Leer nombres de variables*.

En *Vista de variables*:

1. Etiquetar las siguientes variables.

Variable	*Etiqueta*
Num	Número de Encuesta
Her	Número de hermanos
Alt	Estatura
Como	Comodidad

2. Etiquetar los valores de las siguientes variables.

Variable	*Valor*	*Etiqueta*
Genero	6	Hombre
	9	Mujer
Resid	1	Si
	2	No

3. Modificar el tipo de variable.

Variable	*Tipo*	*Ancho*	*Decimales*
Curso	Numérico	2	0
Fecha	Fecha	dd-mmm-aaaa	
Coste	Punto	10	2

5. Establecer un ancho de columna en *Columnas* para la variable Curso igual a 2 y para la variable Coste igual a 10.

6. Modificar los siguientes valores en *Vista de datos*.

Variable	Buscar	Nuevo
Rapi	12	2
Inde	-1	3

Se sitúa el cursor sobre la columna de la variable que se quiere modificar (Rapi). Con la opción *Edición* > *Buscar* o el icono *Buscar* se desplaza el cursor hasta la celda correspondiente al valor (12) y se teclea el nuevo valor (2).

Se procede de la misma forma con la variable Inde. Se sitúa el cursor sobre la columna de esta variable. Con la opción *Edición* >*Buscar* o el icono Buscar se desplaza el cursor hasta la celda correspondiente al valor (-1) y se teclea el nuevo valor (3).

7. Guardar el archivo de datos con formato SPSS extensión SAV y nombre **Ejemplo**.

IMPORTAR UN ARCHIVO EXCEL ODBC

Si los datos están almacenados en una hoja de cálculo Excel de grandes dimensiones o de la versión 5.0 o posteriores, se puede importar al programa SPSS con las opciones del menú del editor de datos:

Archivo

.....

Abrir base de datos

.

Nueva consulta

En la ventana que aparece se debe seleccionar *Excel Files* y, al pulsar *Siguiente*, se pasa al cuadro de diálogo *Seleccionar libro* donde se debe indicar unidad, directorio y nombre de la base de datos. Al *Aceptar* aparece la lista de hojas y rangos con nombre disponibles para leer.

PREGUNTAS

- **¿Cuáles son los componentes de un editor de datos?**
- **¿Cómo se crea, abre y guarda un archivo de datos?**
- **¿Cómo se importa un archivo de Excel?**

1.2 RECODIFICACIÓN Y PONDERACIÓN DE VARIABLES EN SPSS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo el alumno será capaz de recodificar, recalcular y ponderar nuevas variables.

PONDERACIÓN DE VARIABLES EN SPSS

MODIFICAR UN ARCHIVO

Este programa dispone de un gran número de opciones que permiten calcular nuevas variables a partir de las existentes, modificar las variables existentes en el archivo o bien recodificarlas. Estas opciones pueden aplicarse a todos los casos o sólo a algunos casos seleccionados.

CALCULAR NUEVAS VARIABLES

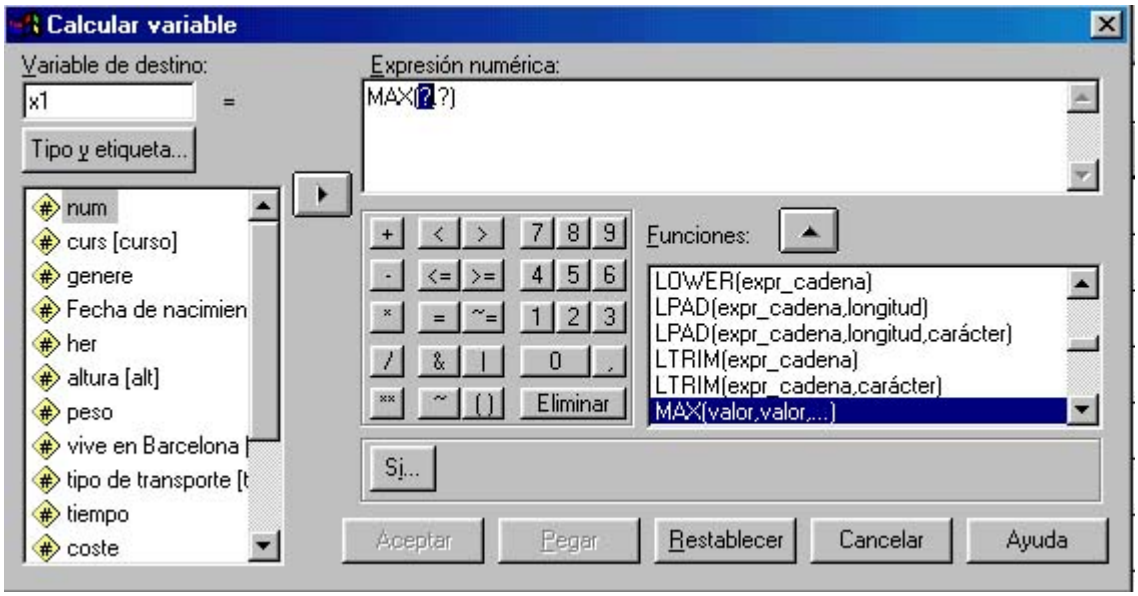
Una vez creado un archivo se pueden generar nuevas variables mediante transformaciones numéricas de las variables existentes.

La opción de menú que se debe activar es:

Transformar

..... *Calcular*

Esta opción abre el cuadro de diálogo *Calcular variable*.



En este cuadro se debe indicar:

• En el recuadro *Variable de destino* se debe asignar un nombre a la nueva variable. Por defecto, el tipo de variable será numérico; para modificarlo, o para añadirle etiquetas, se situará el cursor en el botón *Tipo y etiqueta*.

• En el recuadro *Expresión numérica* se introduce la fórmula a partir de la cual se calculan los valores de la nueva variable. Para escribirla se puede optar entre teclearla directamente o utilizar los botones que aparecen en el cuadro de diálogo.

Las fórmulas pueden contener: nombres de variables existentes, constantes, operadores y funciones.

Operadores: Los operadores disponibles son de tres tipos: aritméticos, relacionales y lógicos. Además de las 4 operaciones aritméticas básicas, los operadores de uso más frecuente son:

**		Exponente
<	LT	Menor que
>	GT	Mayor que
<=	LE	Menor o igual que
>=	GE	Mayor o igual que
=	EQ	Igual a
~ =	NE	Distinto de
&	AND	Y
|	OR	O
~	NOT	No

Funciones: El SPSS contiene más de setenta funciones (véase anexo funciones) entre las que se pueden destacar:

SUM(expr_num,expr_num[,...])

MEAN(expr_num,expr_num[,...])

SD(expr_num,expr_num[,...])

Una vez escrita la fórmula apretando el botón *Aceptar* se genera la variable que queda recogida en el editor de datos. Para salvarla se deberá guardar el archivo de datos antes de finalizar la sesión de trabajo.

CALCULAR NUEVAS VARIABLES PARA CASOS SELECCIONADOS

Cuando la expresión numérica de cálculo de la nueva variable sólo se quiere que afecte a los casos que cumplen una determinada condición, ésta debe especificarse en el cuadro de diálogo que aparece al pulsar el botón si los casos *Si*...

Para introducir la condición se activa, en primer lugar, *Incluir si el caso satisface la condición*. Seguidamente se escribe la expresión condicional utilizando los nombres de variables, operadores y funciones. Al aceptar la condición y la expresión de cálculo de la nueva variable, se calcularán los nuevos valores para los individuos o elementos que cumplen la condición, permaneciendo el resto inalterado.

MODIFICAR UNA VARIABLE

Cuando se desee modificar los valores de una variable (por ejemplo, cambiar las unidades de medida) sin cambiarle el nombre se utilizarán las mismas opciones que en el apartado anterior, pero en el recuadro *Variable de destino* del cuadro de diálogo *Calcular variable* se indica el nombre de la variable original.

EJEMPLO

1. Crear una nueva variable con nombre Media que sea el promedio de las variables Como, Rapi e Inde utilizando operadores (es decir, sin utilizar la función estadística).

La expresión de cálculo es: Media=(Como+Rapi+Inde)/3

Para calcular el promedio de las puntuaciones recogidas en las variables Como, Rapi e Inde se activa la opción de menú Transformar >Calcular. Se indica el nombre de la

nueva variable, Media, en Variable de destino; se introduce la fórmula tecleándola, o bien seleccionando el nombre de las variables del directorio, en Expresión numérica.

2. Crear una nueva variable con nombre Media2 que sea el promedio de las variables Como, Rapi e Inde utilizando la función MEAN.

La función es: Media2=MEAN(Como,Rapi,Inde)

Con la secuencia Transformar >Calcular se abre el cuadro de diálogo. Se selecciona la función MEAN(?,?) y se pega apretando el botón . Por último, se sustituyen los interrogantes que aparecen en la función por el nombre de las variables.

La diferencia entre estos procedimientos se debe a que en el primero el cálculo se realiza sólo si no hay valores missing en ninguna de las tres variables, mientras que el segundo promedia todos los casos, aún cuando presenten algún valor perdido; en consecuencia siempre resulta más eficaz emplear las funciones del sistema.

3. Calcular la edad en años a partir de la fecha de nacimiento.

La expresión para calcular la edad es:

Edad=YYYY * -XDATE.YEAR(Fecha)

En el cuadro de diálogo Calcular variable se introduce el nombre de la nueva variable (Edad) y en la casilla Expresión numérica se escribe la fórmula seleccionando la función XDATE.YEAR (?) y\ sustituyendo el interrogante por la variable Fecha.

4. Cambiar las unidades de medida de la variable Altura de metros a centímetros para aquellos casos en que la variable Curso sea 2.

La expresión que permite calcular la altura en centímetros es:

Alt=Alt*100

En el cuadro de diálogo Calcular variable se da nombre a la variable (Alt) y en la casilla Expresión numérica se escribe la fórmula anterior. Seguidamente, se pulsa el botón Si.... En el nuevo cuadro de diálogo, se selecciona Incluir si el caso satisface la condición y se introduce la condición: Curso=2.

RECODIFICAR VALORES

Cuando se trabaja con variables codificadas y se quiere recodificar sus valores, ya sea porque la codificación inicial se revela inadecuada, o porque se desea unificar criterios de recodificación, o porque interesa codificarlas o definir intervalos de valores (categorizar variables continuas o combinar varias categorías discretas en una sola) puede procederse de dos maneras:

• Recodificar en la misma variable de forma que los valores nuevos sustituyan a los antiguos.

• Generar una nueva variable cuyos valores son el resultado de recodificar los antiguos.

Recodificar en la misma variable:

Para recodificar en la misma variable la secuencia es:

Transformar
Recodificar
En las mismas variables

Se abre el cuadro de diálogo *Recodificar en las mismas variables*. En primer lugar, en la ventana *Variables numéricas* se eligen las variables que se desea recodificar. Puede recodificarse todos los casos o solamente algunos casos seleccionados. Si sólo se quieren recodificar los casos que verifiquen una condición, se apreta el botón *Si.*. y se accede a la ventana *Recodificar si los casos* cuyo funcionamiento ya ha sido explicado en el apartado anterior.

El botón *Valores antiguos y nuevos* abre el cuadro de diálogo que permite especificar el criterio de codificación:

Para introducir las equivalencias del criterio de codificación deben indicarse los *Valores antiguos*, que pueden ser:

- un valor concreto, entonces debe seleccionarse Valor;

- un intervalo de valores, entonces debe elegirse Rango que puede estar entre dos valores concretos, desde el menor hasta un valor concreto o desde un valor concreto hasta el mayor.

En el cuadro de la derecha (*Valor nuevo*) debe especificarse el nuevo valor que sustituirá a los anteriores (valor concreto, valor missing).

Por último, pulsando el botón *Añadir* se traslada la equivalencia al cuadro resumen *Antiguo --> Nuevo*. Si en cualquier momento del proceso se desea modificar o borrar alguna equivalencia es posible hacerlo con los botones *Cambiar y Borrar*, respectivamente.

Una vez introducidas todas las equivalencias con el botón *Continuar* se vuelve al cuadro de diálogo *Recodificar en las mismas variable* y pulsando el botón *Aceptar* se ejecuta la transformación.

Recodificar en distintas variables:

Para recodificar en distintas variables la secuencia es:

Transformar
Recodificar
En distintas variables

Se abre el siguiente cuadro de diálogo cuya estructura es muy similar al anterior, la única diferencia es que contiene un cuadro de texto donde se debe especificar el nombre de la variable, y de la etiqueta si se desea, que se va a generar.

EJEMPLO

1. Recodificar los valores perdidos (missing) de la variable Her en la misma variable aplicando la siguiente equivalencia:

Valor antiguo
Valor nuevo . - 9
y etiquetar el nuevo valor con 'No contesta'.

2. Recodificar la variable Coste en una nueva variable de nombre Coste2 que contemple tres categorías:

Valor antiguo	Valor nuevo
de 0 a 5000	1
de 5000 a 10000	2
más de 10000	3

Etiquetar los valores de esta nueva variable: 1 'Bajo', 2 'Medio' y 3 'Alto'.

3. Crear una nueva variable con nombre Valora que recoja la característica más valorada por el usuario entre Como, Rapi e Inde, asignando los valores 1 si es comodidad, 2 si es rapidez y 3 si es independencia.

Entre las posibles opciones para resolver esta cuestión puede utilizarse la siguiente:

a. Recodificar la variable Como en una nueva variable de nombre Valora siguiendo los siguientes criterios de recodificación:

Condición Si
Valor antiguo
Valor nuevo

Como=Max(Como,Rapi,Inde)
Todos los demás valores 1

Con esta expresión la nueva variable toma el valor 1 en aquellos individuos que consideran la comodidad como la característica más valorada y missing el resto de valores.

b. Asignar el valor 2 a la variable Valora en aquellos usuarios que consideran la rapidez como la característica más valorada. Para ello debe recodificarse en la misma variable Valora aplicando el siguiente criterio:

Condición Si
Valor antiguo
Valor nuevo

Rapi=Max(Como,Rapi,Inde)
Todos los demás valores 2

c. Asignar el valor 3 en la variable Valora a los usuarios que consideran la independencia por encima de las otras características. Para lo cual debe procederse de la siguiente manera:

Condición Si
Valor antiguo
Valor nuevo

Inde=Max(Como,Rapi,Inde)
Todos los demás valores 3

d. Etiquetar los valores de la variable Valora como: 1 'Comodidad', 2 'Rapidez' y 3 'Independencia'.

SELECCIÓN Y PONDERACIÓN DE CASOS

Las modificaciones de un archivo de datos pueden referirse, no sólo a las variables, sino también a los casos o elementos con el objeto de seleccionar o filtrar algunos de ellos, o bien, de asignar a cada caso una ponderación que indique el peso relativo del mismo dentro del conjunto.

SELECCIÓN DE CASOS

En ocasiones de todos los casos que integran un archivo interesa seleccionar algunos de ellos, ya sea aleatoriamente para obtener una muestra reducida o bien de acuerdo con una determinada condición referida a una o más variables.

La selección de casos puede realizarse desde la barra de menús con la secuencia:

Opción de menú:

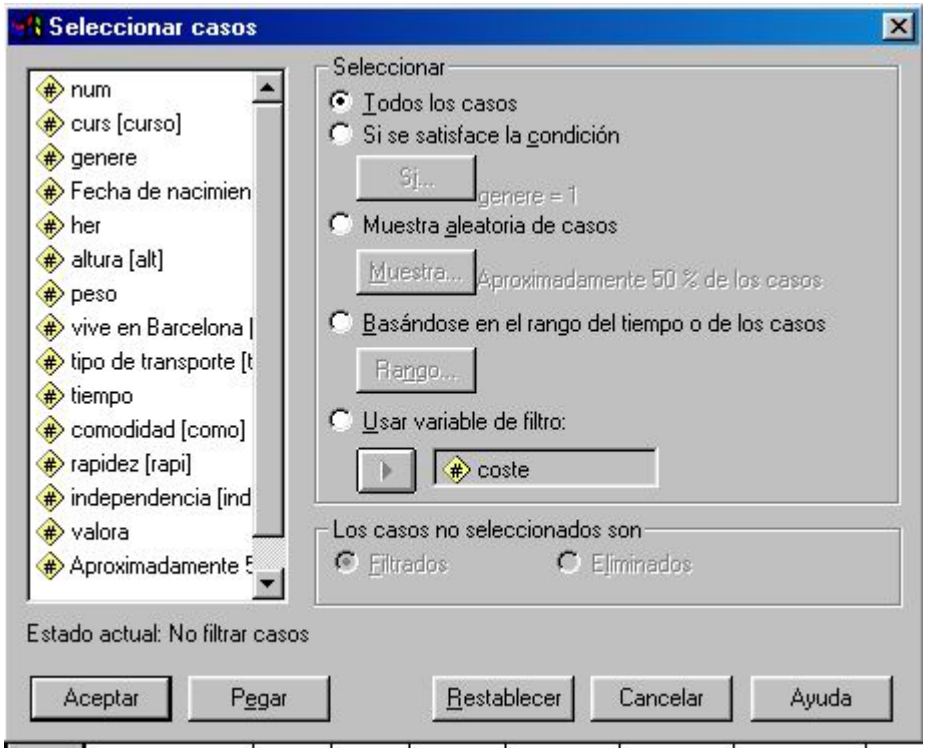
Datos

Seleccionar casos

o activando la barra de herramientas del icono: 

La opción activada por defecto en el cuadro de diálogo *Seleccionar casos* es *Todos los casos* y bajo la lista de variables aparece el mensaje *Estado actual: No filtrar casos*.

Antes de proceder a la selección de casos hay que tener en cuenta que el programa presenta dos formas distintas para el tratamiento de los casos no seleccionados dependiendo de la opción que se active en el recuadro *Los casos no seleccionados son*. Con la opción *Filtrados* los casos no seleccionados no son incluidos en ningún procedimiento posterior pero permanecen en la base de datos y pueden recuperarse en cualquier momento desactivando el filtro; con la opción *Eliminados* los casos no seleccionados desaparecen de la base de datos y ya no son recuperables.



La selección de datos puede realizarse con:

1. *Si se satisface la condición.* Aparece el cuadro de diálogo *Seleccionar casos: Si* en donde debe especificarse la condición de filtrado, de manera que los casos que la verifiquen quedan seleccionados.
2. *Muestra aleatoria de casos.* Con esta opción se selecciona una muestra aleatoria fijando el tamaño de la misma como un porcentaje del total o como un número exacto de casos.
3. Basándose en el rango del tiempo o de los casos. Esta opción es únicamente válida para datos de series temporales, basando la selección en intervalos de tiempo.
4. Usar variable de filtro. En este caso debe elegirse una variable numérica del archivo, y ésta actúa como filtro en el sentido de que quedan seleccionados aquellos casos que en la variable filtro toman valores distintos de cero y no son missing.

Siempre que hay un filtro activado a la derecha de la barra de estado aparece la palabra Filtrado. Los casos no seleccionados aparecen marcados en la base de datos.

PONDERAR CASOS

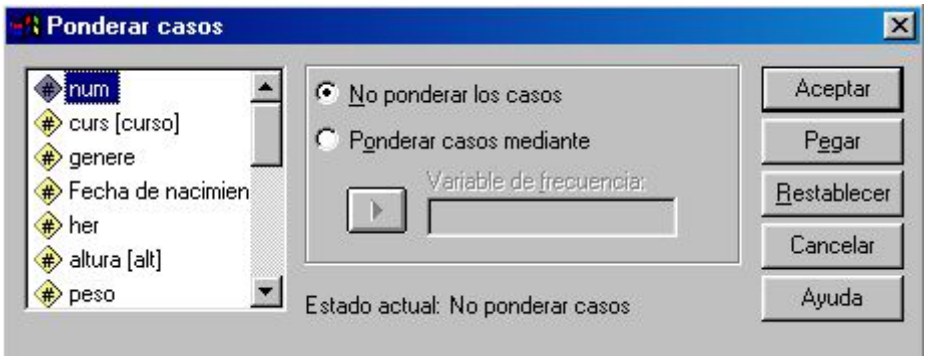
Algunos análisis estadísticos requieren emplear coeficientes de ponderación para asignar importancias diferentes a los valores de la variable. El SPSS permite utilizar una variable como criterio de ponderación, el único requisito es que dicha variable esté en el archivo activo y sus valores serán las ponderaciones.

El cuadro de diálogo *Ponderar casos* puede abrirse con la secuencia:

Datos

Ponderar casos

o activando el icono  de la barra de herramientas



Se establece la ponderación activando *Ponderar casos mediante* y seleccionando la *Variable de frecuencia* que recoge los coeficientes de ponderación de la lista de variables. A partir de este momento todos los análisis se realizarán teniendo en cuenta estas ponderaciones. Para desactivar la ponderación se elige del cuadro de diálogo anterior la opción *No ponderar los casos*.

La barra de estado del editor de datos nos indica si hay ponderación activada.

EJEMPLO

En un archivo en el que *figura la siguiente información acerca de la inversión realizada en títulos de renta variable por un inversor.*

Las variables son:

Nombre- Nombre del título.

Precio - Precio en Euros de un título al cierre de la sesión de hoy.

Variac - Variación del precio en Euros respecto a la sesión anterior.

Numero - Número de títulos que posee el inversior.

Si todos los títulos los adquirió a precio de cierre de la sesión de hoy, determine la cantidad total invertida, en Euros y la cantidad media invertida por título.

Se abre el archivo de datos

Se desplaza la variable cuyos valores son las ponderaciones, en este caso Numero, a la casilla *Variable de frecuencia* del cuadro de diálogo que se abre con la secuencia *Datos > Ponderar casos*, o bien con el icono *Ponderar*.

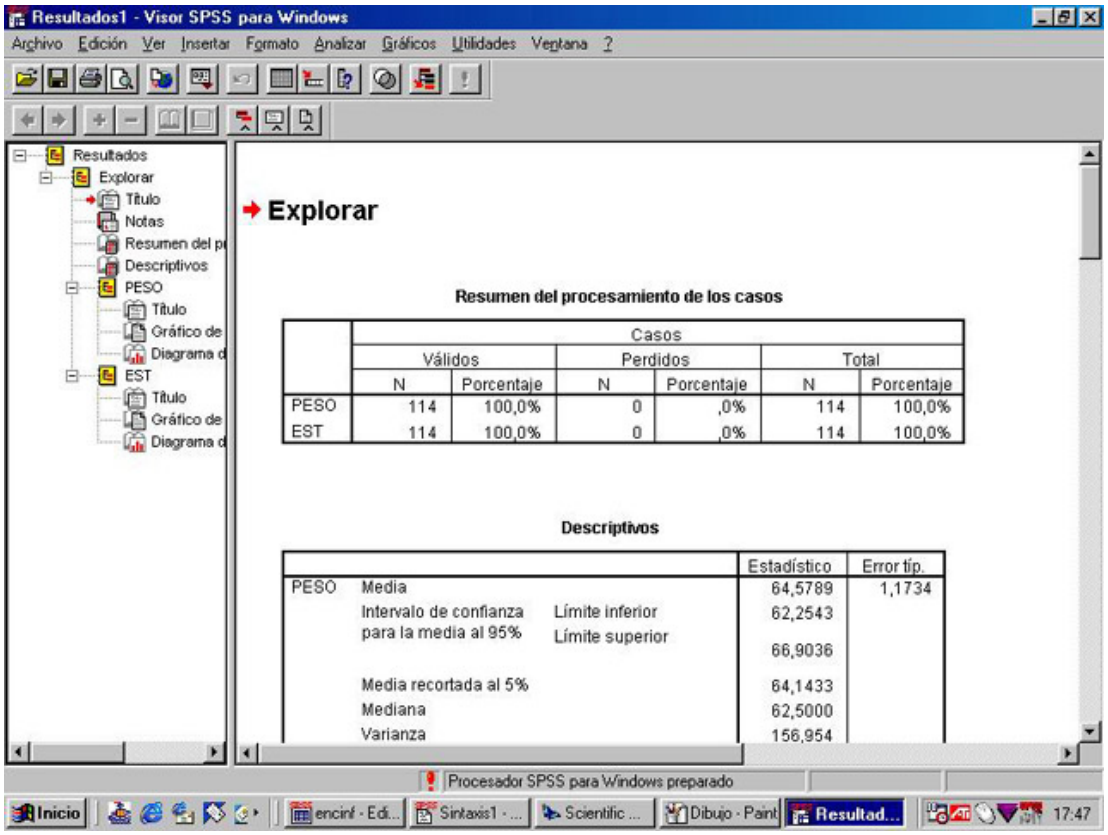
Con la secuencia *Analizar* > *Estadísticos Descriptivos* > *Descriptivos* se abre el cuadro de diálogo donde se selecciona la variable Precio y en *Opciones* se activa *Media y Suma*. Los resultados que se obtienen son:

Estadísticos descriptivos

	N	Suma	Media
PRECIO	290	5172,25	17,8353
N válido (según lista)	290		

EDITOR DE RESULTADOS

Por defecto el programa va almacenando los resultado en un documento con extensión (*.spo) que se deberá salvar antes de finalizar la sesión si se quiere conservar. Este documento admite modificaciones y se visualiza en el visor de resultados o output.



La ventana del visor de resultados está dividida en dos paneles:

1. Panel de titulares. Se encuentra en la parte izquierda de la ventana y recoge un índice de todos los resultados generados durante la sesión. Este índice facilita el desplazamiento de un resultado a otro; permite ocultar sin eleminar, borrar o

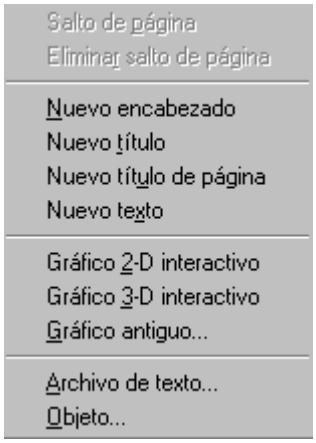
seleccionar grupos de resultados; también permite modificar el orden de los resultados*.

2. Panel de contenidos. Está situado en la parte derecha de la ventana y contiene los resultados: tablas estadísticas, gráficos y comentarios. Este panel permite modificar el contenido y formato del documento. Por ejemplo, añadir nuevos elementos: encabezado, títulos, textos; o editar los elementos ya existentes para modificarlos.

En la parte superior del visor de resultados hay las siguientes barras:

1. Barra de *menús*. Básicamente recoge los mismos menús que el editor de datos pero en algún caso las opciones son diferentes. Los menús específicos del visor de resultados son:

• *Insertar*. Este menú permite incorporar nuevos elementos en el documento de resultados.



• *Formato*. Permite modificar la alineación de los resultados: títulos, valores, etiquetas,...

2. *Barra de herramientas.* Algunas herramientas son especificas del visor y recogen opciones que modifican la visualización y edición de los resultados o insertan elementos al documento.

•  Desplaza el cursor al editor de datos.

•  Modifican la visualización del panel de titulares. Los dos primeros cambian el nivel de los elementos (degradar \leftarrow o ascender \rightarrow), los dos centrales la visualización de los elementos (expandir + o contraer -) y los dos últimos ocultan elementos

(libro abierto-mostrar, libro cerrado-ocultar).

•  Incorporan elementos en el panel de contenidos. El primero Inserta un encabezado, el segundo un título y el tercero un texto (equivalen a estas opciones del menú insertar).

En la parte inferior de la ventana se encuentra la línea de mensajes. Informa del estado actual del editor y de la operación que se debe realizar para acceder o editar el elemento seleccionado.

EDICIÓN DE RESULTADOS EN EL VISOR

Si se quiere modificar las características de algún elemento del panel de contenidos, éste se deberá editar. Para ello se sitúa el cursor sobre el elemento (texto, tabla o gráfico) y se pulsa dos veces el botón izquierdo del ratón. La pantalla que se abre depende del tipo de elemento que se está editando:

1. *Texto*. Si el elemento editado es un texto podemos modificar sus características, fuentes, colores, alineación con el menú *Formato*. Con las opciones del menú *Edición* podemos copiar, mover o eliminar el texto seleccionado.

2. *Tabla*. Al editar una tabla, la barra de menús que aparece presenta, entre otros, los menús:

• *Insertar*. Permite insertar un título, un texto al pie, es decir, en la parte inferior de la tabla o una nota al pie de la página.

• *Formato*. Incluye opciones que modifican el aspecto y las propiedades de la tabla. Los más importantes son:

Propiedades de la casilla. Permite modificar el formato de los valores, la alineación (horizontal, vertical), los márgenes y el sombreado de las celdas seleccionadas.

Propiedades de la tabla. Afecta al formato de toda la tabla. El menú *General* modifica el ancho de las columnas y de las filas. En el menú Formato de casillas se selecciona la fuente, el tamaño, la alineación, el sombreado, colores y los márgenes internos.

Aspectos de la tabla. Proporciona una amplia gama de diseños de tablas con diferentes bordes y espaciados. También permite elegir un aspecto personalizado por defecto o modificar el formato de tabla que aparece por defecto. Para ello, se debe Editar el aspecto, cambiar las propiedades que se deseen y guardarlo pulsando la opción *Guardar aspecto* o *Guardar como*.

GUARDAR Y EXPORTAR RESULTADOS

GUARDAR

Antes de finalizar la sesión, los documentos de resultados que se quieran conservar deben salvarse.

Para salvar un archivo se elige en el menú del visor de resultados:

Archivo
Guardar como

Si el documento es nuevo se abre un cuadro de diálogo donde se debe indicar la unidad, el nombre y el formato (por defecto *.spo). Si existía al iniciar la sesión se almacena automáticamente. En este último caso también se puede guardar modificando alguna de sus características con las opciones:

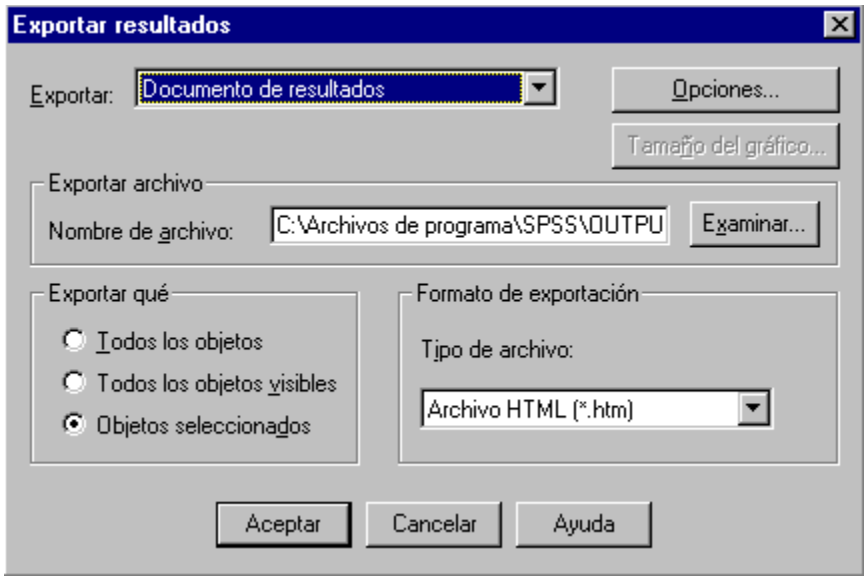
Archivo
Guardar como

EXPORTAR

Los archivos de resultados SPSS no pueden abrirse con otras aplicaciones (por ejemplo, Word, Excel,...). Esto obliga a tener que exportarlos para incorporarlos en otros documentos ya que la solución Copiar y Pegar o Pegar como (en otra aplicación) puede generar problemas si se realiza para todo un bloque de resultados.

Para exportar se debe elegir en el menú del visor de resultados:

Archivo
Exportar



En el cuadro de diálogo Exportar se debe indicar:

• En la lista desplegable Exportar, el tipo de documento que puede ser:

Documento de resultados
Documento de resultados sin gráficos
Sólo gráficos

• En la lista Nombre de archivo que se abre con el botón Examinar, la carpeta o subdirectorio y el nombre del archivo.

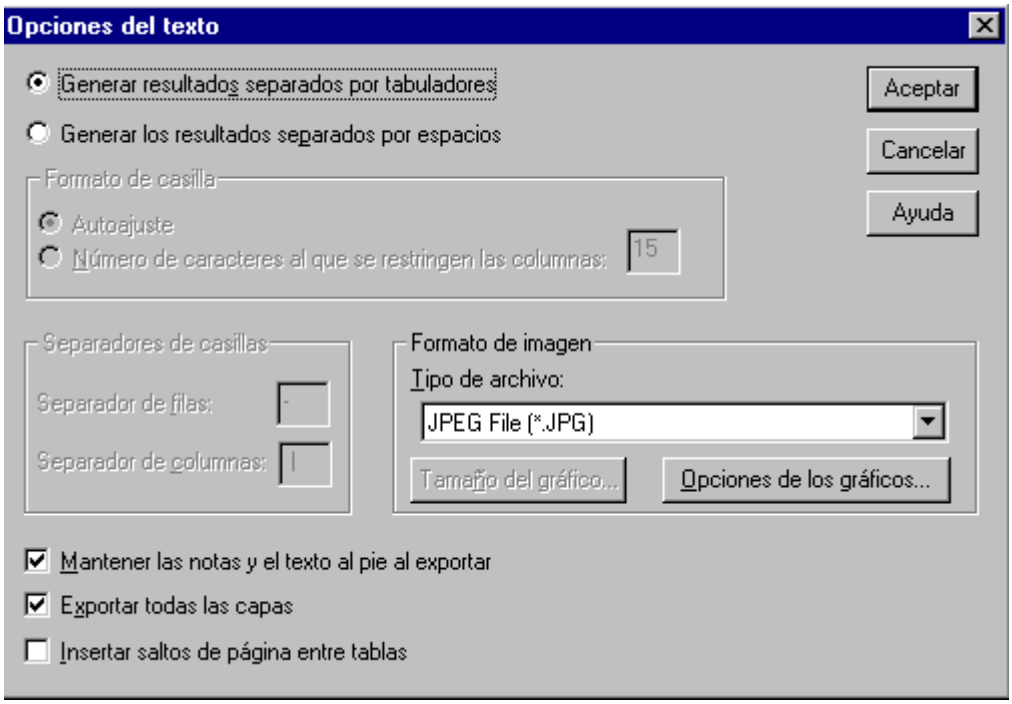
• En Exportar qué, la parte del documento que se quiere exportar, pudiendo elegir entre:

Todos los objetos
Todos los objetos visibles
Objetos seleccionados

• En Formato de exportación, el tipo de archivo entre los contenidos en la lista Tipo de archivo. éste dependerá del tipo de documento que se exporta. Si es texto y gráficos o sólo texto la opciones son: archivo HTML o archivo texto. Los gráficos pueden exportarse en formato JPEG file, CGM Metafile, Windows Bitmap, Windows Metafile, entre otros.

El botón Opciones abre un cuadro de diálogo que depende del tipo del documento que se

exporta. Así, por ejemplo, si el documento es e resultados y gráficos y el formato de exportación es tipo texto se obtiene el cuadro:



Es muy importante indicar el tipo de separador (tabuladores o espacios) y el tipo de archivo de los gráficos.

DISTRIBUCIONES UNIDIMENSIONALES: INTRODUCCIÓN

Todo análisis estadístico se inicia con una primera fase descriptiva de los datos. ésta tiene por objeto sintetizar la información mediante la elaboración de tablas de frecuencias, representaciones gráficas y el cálculo de medidas estadísticas (o estadísticos). Estos procedimientos descriptivos dependen de la naturaleza de la variable o atributo que se analiza y, en este sentido, el programa SPSS los recoge en dos menús diferentes según se empleen, básicamente, para sintetizar datos cualitativos o datos cuantitativos. Así mismo, el programa diferencia entre los procedimientos descriptivos que hacen referencia al análisis de una sola variable (análisis unidimensional) de los relativos a dos o más variables conjuntamente (análisis bidimensional o multidimensional).

PREGUNTAS

- **¿Cómo se realiza una recodificación de variables?**
- **¿Cómo se realiza una ponderación de variables?**

Segunda unidad

……

Estadística Descriptiva

2.1 FRECUENCIAS Y GRÁFICOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo el alumno podrá realizar correctamente el reporte de resultados de frecuencias y la elaboración de gráficos básicos.

DISTRIBUCIONES DE FRECUENCIAS

Las distribuciones o tablas de frecuencias permiten resumir los datos en una tabla que recoge:

• *valores* de la variable o modalidades del atributo,

• *frecuencia absoluta* o número de veces que aparece cada valor o modalidad en la muestra,

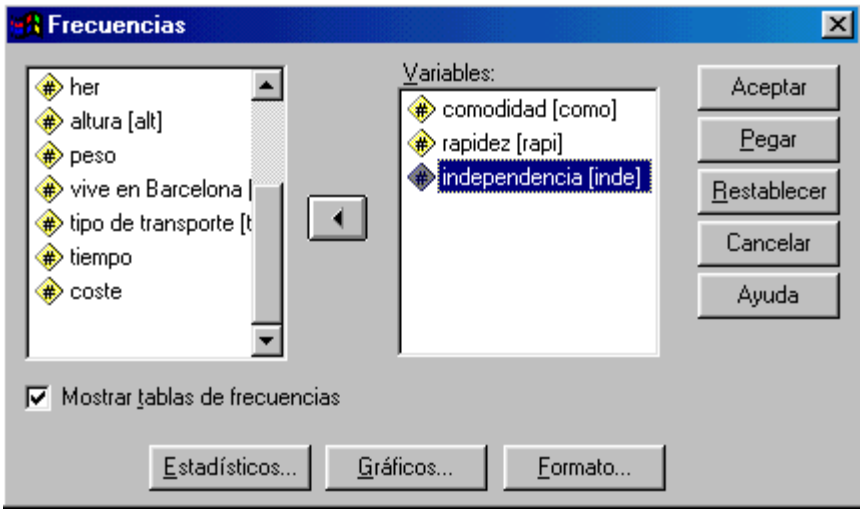
• *porcentaje* de veces que aparece cada valor de la variable o modalidad del atributo sobre el total de observaciones,

• *porcentaje válido* calculado sobre el total de observaciones excluidos los valores missing,

• *porcentaje acumulado* hasta cada uno de los valores de la variable ordenados de menor a mayor. Este porcentaje tiene interpretación sólo en los casos en que la variable sea susceptible de medida por lo menos en una escala ordinal.

Para obtener la tabla de frecuencias se procede con el menú:

Analizar
Estadísticos Descriptivos
Frecuencias



En el cuadro de diálogo *Frecuencias* se seleccionan las variables para las que se quiere obtener sus correspondientes tablas de frecuencias unidimensional y se trasladan al cuadro *Variables* con el botón  . Para obtener la distribución de frecuencias debe estar activada la opción *Mostrar tablas de frecuencias*. La tabla que aparece en el visor de resultados no agrupa en intervalos o clases los valores de la variable; si se desea agruparlos es necesario recodificar previamente la variable (en otra variable) definiendo los límites de los intervalos*

Además, el cuadro de diálogo *Frecuencias* permite activar otras opciones con los botones:

• *Estadísticos*

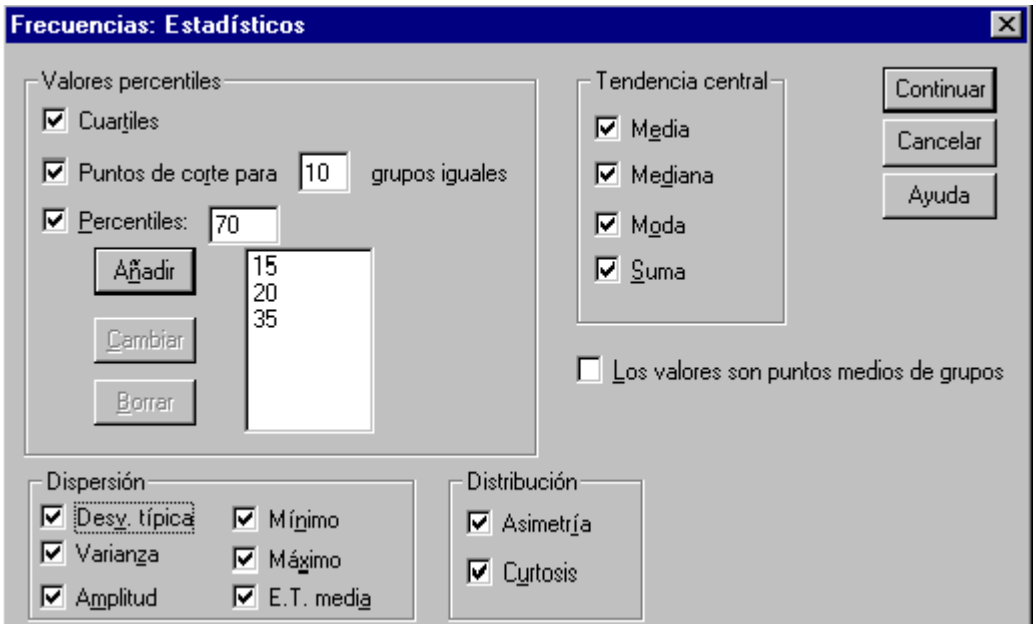
• *Gráficos*

• *Formato*

Estas opciones pueden utilizarse teniendo o no activada la opción *Mostrar tablas de frecuencias*.

ESTADÍSTICOS

La opción *Estadísticos* abre un cuadro de diálogo que permite la obtención de las principales medidas de síntesis o estadísticos de una distribución unidimensional de frecuencias. éstos se presentan agrupados en cuatro clases: *Valores percentiles, Tendencia central, Dispersión y Distribución*.



• *Valores percentiles* son aquellos valores de la variable que dividen a la distribución de frecuencias en partes con igual número de observaciones: así, los cuartiles la dividen en cuatro partes guales y se obtienen directamente activando la opción *Cuartiles*. Si interesan los valores que dividen la distribución en k partes iguales se activa la opción *Puntos de corte para* (por defecto 10) *grupos iguales*, lo que proporciona los deciles de la distribución. En la opción *Percentiles* es necesario indicar cuales de ellos se desean, incluyéndolos de uno en uno con el botón *Añadir*.

• *Tendencia central* permite seleccionar *Media, Mediana y Moda* de la distribución, así como la *Suma* de todos los valores de la distribución.

• *Dispersión* permite seleccionar las siguientes medidas: *Varianza*, como resultado del cálculo de la expresión, $s^2 = \frac{\sum_{i=1}^{k}(X_i - \bar{X})^2 n_i}{n-1}$ y *Desviación típica*; el error típico de la media (*E.T.media*) que se define como $\frac{S}{\sqrt{n}}$, así como los valores *Mínimo* y *Máximo* de la variable y la *Amplitud* o recorrido de la variable.

• Por último, en *Distribución* pueden obtenerse las siguientes medidas relativas a la forma de la distribución:
coeficiente de *Asimetría*, error típico de asimetría, coeficiente de *Curtosis* y error típico de curtosis, calculadas mediantelas siguientes expresiones:

Asimetría $$g_1 = \frac{\frac{nS_3}{(n-1)(n-2)}}{S^3}$$

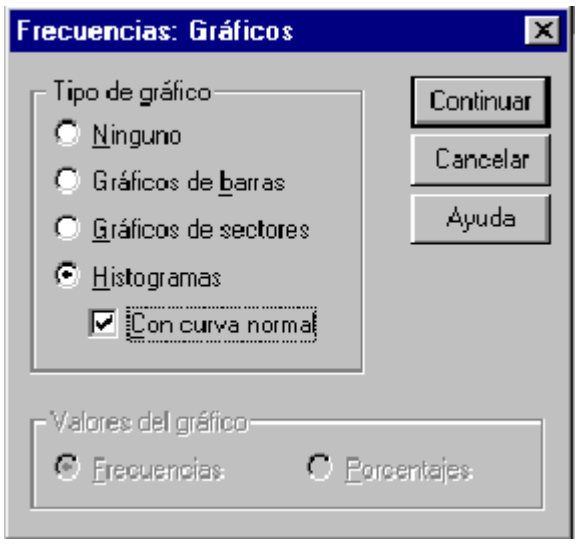
Error típ. de asimetría $$e_{g_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

Curtosis $$g_2 = \frac{\frac{n(n+1)S_4}{(n-1)(n-2)(n-3)} - 3\frac{S^2}{(n-2)(n-3)}}{S^4}$$

Error típ.de curtosis $$e_{g_2} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$$

GRÁFICOS

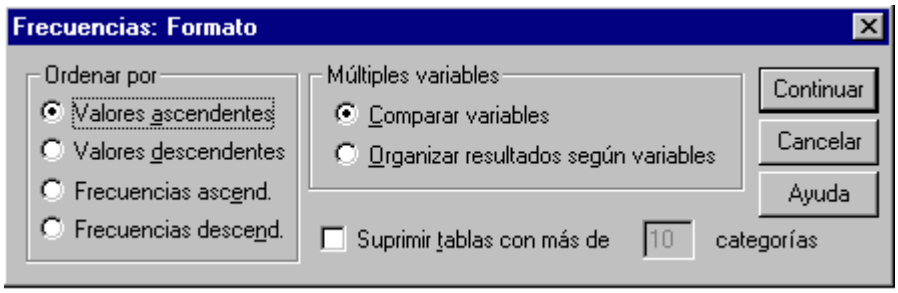
Los gráficos asociados a la tabla de frecuencias que recoge del cuadro de diálogo *Frecuencias* son: *Gráficos de barras*, *Gráficos de sectores* o *Histogramas*. Para seleccionar el que interesa se activa la opción *Gráficos* que abre el siguiente cuadro de diálogo:



Si la característica objeto de análisis es un atributo los gráficos adecuados son el gráfico de barras o de sectores; en ambos casos pueden realizarse con frecuencias absolutas o con relativas seleccionando *Frecuencias* o *Porcentajes*, respectivamente. Si la característica es cuantitativa el gráfico adecuado es el histograma que, a su vez, puede obtenerse superponiéndole la *Curva de la distribución normal* activando la opción correspondiente.

FORMATO

Para modificar el aspecto de los resultados, ya sean, tablas o estadísticos, se activa la opción *Formato* que abre el cuadro de diálogo siguiente:



Con las siguientes opciones:

• *Ordenar por*: se puede elegir entre distintos criterios de ordenación de los valores de la variable en la tabla de frecuencias. Por defecto, los valores aparecen en orden ascendente; pero también es posible una ordenación descendente o una ordenación por frecuencias, tanto ascendente como descendente, activando las opciones correspondientes.

• *Múltiples variables*: se puede seleccionar el tipo de presentación de los cuadros de estadísticos cuando se realiza simultáneamente el análisis unidimensional de dos o más variables. Por defecto, está activada la opción *Comparar variables* que proporciona un único cuadro que contiene los estadísticos seleccionados correspondientes a todas las variables. Si se selecciona la opción *Organizar resultados según variables* se obtiene un cuadro de estadísticos para cada variable por separado.

El cuadro *Frecuencias*: *Formato* también ofrece la posibilidad de limitar la elaboración de tablas de frecuencias sólo para Aquellas variables que presentan un número reducido de valores o categorías. Para ello se debe indicar en el recuadro *Suprimir tablas con más de* (por defecto 10) *categorías* el número de categorías a partir del cual no se desea la elaboración de la tabla.

EJEMPLOS

Ejemplo 1. *Con la base de datos* ***Enctran.sav*** *obtener la tabla de frecuencias, el diagrama de barras y los estadísticos media, mediana, moda, desviación tipo, varianza y las medidas de forma (asimetría y curtosis) de las variables: Como, Rapi e Inde.*

Vamos a realizar la descripción de la variable Como, dejando al lector la descripción de las variables Rapi e Inde.

Con la secuencia *Analizar* **>** *Estadísticos Descriptivos* **>** *Frecuencias* se abre un cuadro de diálogo donde se selecciona la variable Como; con el botón *Estadísticos* se activan las medidas que se desean obtener y con el botón *Gráficos* se activa la opción *Gráficos de barras*.

Se obtienen los siguientes cuadros:

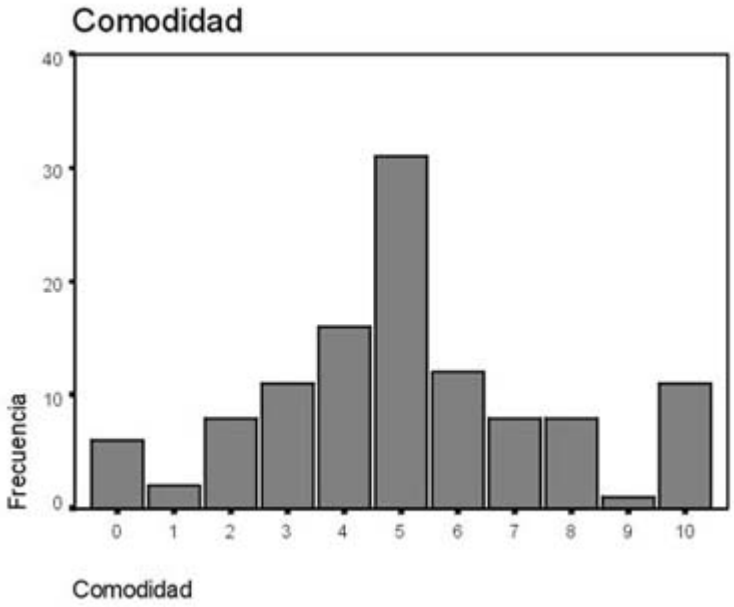
Estadísticos

COMO Comodidad

N	Válidos	114
	Perdidos	0
Media		5,10
Mediana		5,00
Moda		5
Desv. típ.		2,52
Varianza		6,37
Asimetría		,182
Error típ. de asimetría		,226
Curtosis		-,126
Error típ. de curtosis		,449

COMO Comodidad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	6	5,3	5,3	5,3
	1	2	1,8	1,8	7,0
	2	8	7,0	7,0	14,0
	3	11	9,6	9,6	23,7
	4	16	14,0	14,0	37,7
	5	31	27,2	27,2	64,9
	6	12	10,5	10,5	75,4
	7	8	7,0	7,0	82,5
	8	8	7,0	7,0	89,5
	9	1	,9	,9	90,4
	10	11	9,6	9,6	100,0
	Total	114	100,0	100,0	



En base a estos resultados se concluye:

- La base de datos no presenta para esta variable ningún valor missing, de forma que las 114 observaciones son todas válidas.

- Las medidas de posición (media, mediana y moda) indican el valor central de la distribución, y en este caso aproximadamente coinciden los tres estadísticos en el valor 5. Esto significa que la distribución es bastante simétrica y que la valoración media de la comodidad del medio de transporte no es ni buena ni mala.

- La desviación típica es 2,52 que sobre una media de 5,1 indica que la dispersión de los datos con respecto a la media es moderada.

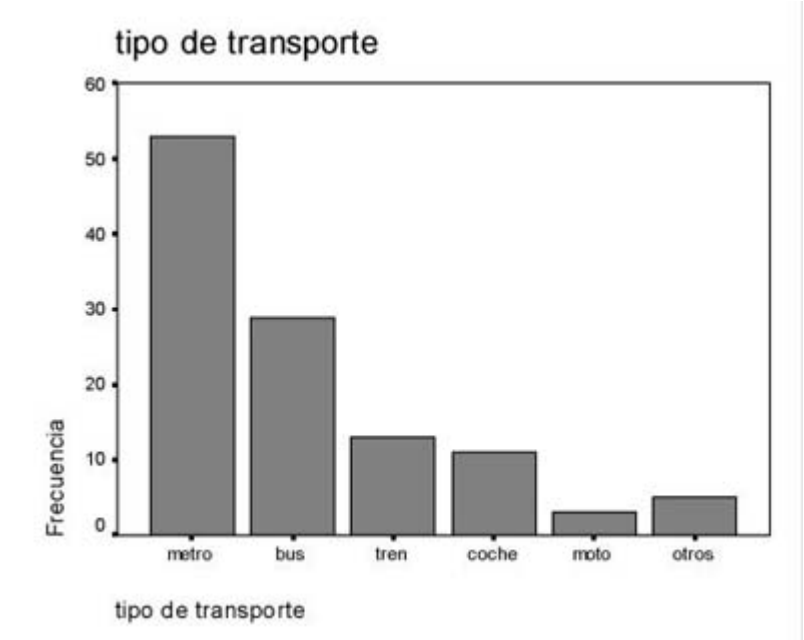
- El coeficiente de asimetría toma el valor 0,182 que no es significativo ya que presenta un error estándar 0,226 y, por lo tanto, puede considerarse que la distribución es simétrica. La curtosis de esta variable es -0,126 con un error estándar de 0,449 lo que indica que la distribución es mesocúrtica.

- La distribución de la variable es unimodal, prácticamente simétrica y campaniforme como se observa en el gráfico.

Ejemplo 2. *Con la misma base de datos **Enctran.sav** obtener la tabla de frecuencias y el diagrama de barras de la variable Trans*.

TRANS tipo de transporte

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1 metro	53	46,5	46,5	46,5
	2 bus	29	25,4	25,4	71,9
	3 tren	13	11,4	11,4	83,3
	4 coche	11	9,6	9,6	93,0
	5 moto	3	2,6	2,6	95,6
	7 otros	5	4,4	4,4	100,0
	Total	114	100,0	100,0	



Entre otros resultados se observa que los porcentajes correspondientes a las tres modalidades de transporte público acumulan el 83,3% de los estudiantes y, únicamente, el 12,2% utiliza transporte privado. Al ser una variable cualitativa el único estadístico representativo de la distribución es la moda que, en este caso, es la modalidad Metro que representa un 46,5% del total.

Ejemplo 3. *Con la base de datos **Enctran.sav** obtener la tabla de frecuencias y la representación gráfica adecuada para la variable Coste, agrupando los valores en los siguientes intervalos: [0,5000) [5000,10000) [10000,15000) [15000, 20000).*

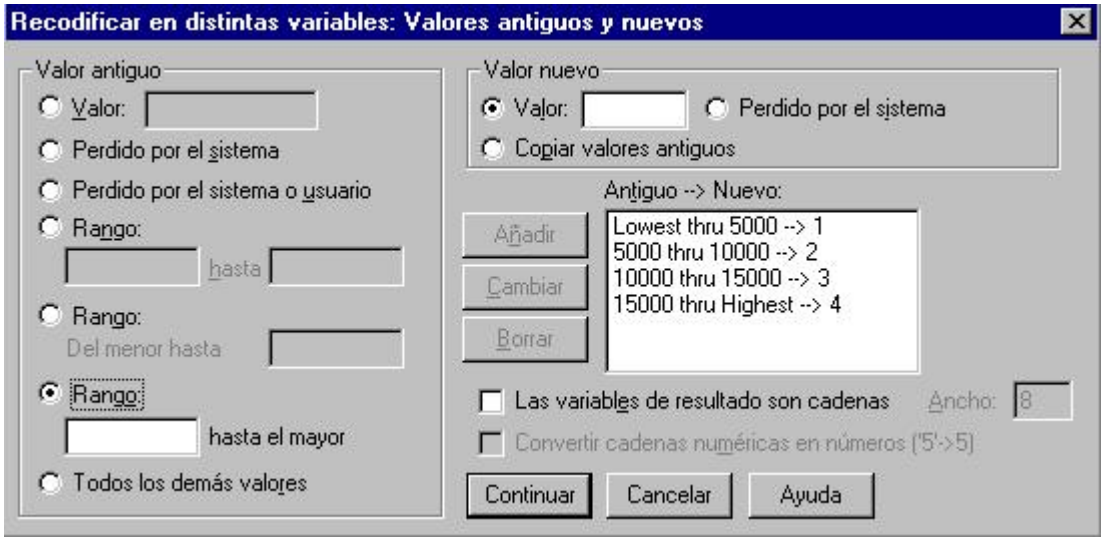
Para obtener la tabla con los valores agrupados en intervalos es necesario, en primer lugar, recodificar los valores en una nueva variable. Para ello, se activa la opción *Recodificar > En distintas variables* del menú *Transformar*. En el cuadro de diálogo que aparece:

• Se selecciona la variable Coste.

• En *Variable de Resultado* se indica el nombre elegido para la nueva variable, por ejemplo, 'Coste1'.

• Se etiqueta la nueva variable, 'Coste recodificado'.

• Se definen los intervalos activando *Valores antiguos y nuevos*. Para definir el primer intervalo se activa en *Valor antiguo* la opción *Rango*: *Del menor hasta 5000* y se le asigna como *Valor nuevo* 1; los siguientes intervalos se definen activando *Rango* límite inferior hasta límite superior, asignándoles los valores 2 y 3. El último intervalo se define mediante *Rango* 15000 *hasta el mayor* y se le asigna *Valor nuevo* 4.



• Se etiquetan los valores de la variable Coste1. En la ventana *Vista de variable* o bien con doble clic sobre la variable Coste1, introducimos las etiquetas de los valores. En *Valores* se indica:

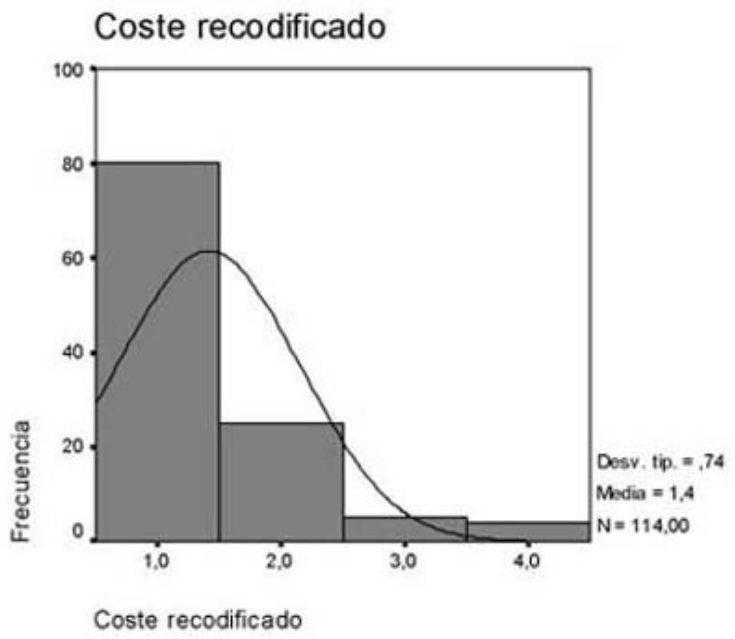
Valor	Etiqueta de

	valor
1	0-5000
2	5000-10000
3	10000-15000
4	15000-20000

• Por último, con *Analizar* > *Estadísticos Descriptivos* > *Frecuencias* se obtiene la tabla de frecuencias y el histograma, que es el adecuado dada la naturaleza continua de la variable.

Coste recodificado

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0-5000	80	70,2	70,2	70,2
	5000-10000	25	21,9	21,9	92,1
	10000-15000	5	4,4	4,4	96,5
	15000-20000	4	3,5	3,5	100,0
	Total	114	100,0	100,0	

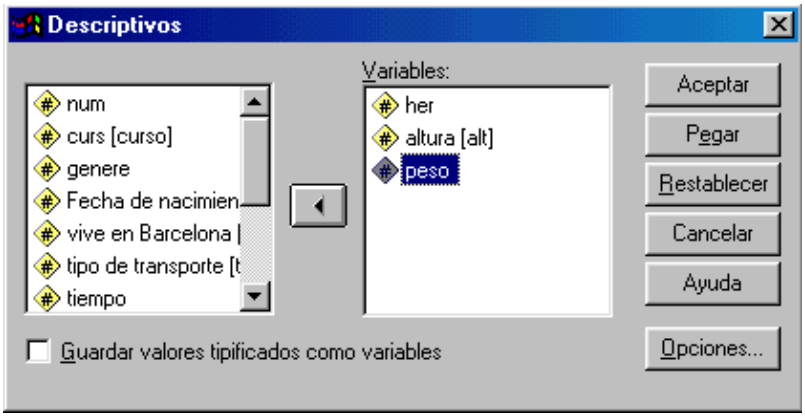


DESCRIPTIVOS

Las principales medidas de síntesis de una distribución pueden obtenerse con la secuencia *Analizar* > *Estadísticos Descriptivos* > *Frecuencias*. El programa contiene además procedimiento *Descriptivos* que está especialmente orientado al análisis y

síntesis de datos cuantitativos (variables continuas). Las opciones de este procedimiento recogen los principales estadísticos univariantes que pueden calcularse para una o varias variables. Para acceder en el menú se elige:

Analizar
Estadísticos
Descriptivos
Descriptivos



En el cuadro de diálogo se debe indicar las variables que se quieren >analizar. Una vez seleccionadas, si se acepta sin modificar las opciones, se obtienen para cada variable: valor mínimo, valor máximo, media y desviación típica.

Para modificar la selección de estadísticos se debe activar *Opciones* antes de aceptar.

Los estadísticos que recoge el cuadro de diálogo Opciones son:

• Medidas de tendencia central: media y suma de todos los valores.

• Medidas de dispersión: desviación típica, varianza, amplitud (o rango), valores mínimo y máximo, E.T. media (error estándar).

• Distribución: coeficientes de curtosis y asimetría.

Por último, el cuadro de diálogo Opciones permite modificar el orden de visualización de las variables analizadas.

El cuadro de diálogo *Descriptivos* permite guardar las puntuaciones estandarizadas (valores tipificados) de las variables seleccionadas. Para ello se activa *Guardar valores tipificados como variables*. Al aceptar, los valores tipificados se añaden en el editor de datos, y quedan disponibles para posteriores análisis. La variable creada toma el nombre Zvar en donde var es el nombre de la variable origen.

El procedimiento *Descriptivos* puede utilizarse con datos cuantitativos agrupados y sin agrupar. En este último caso, previamente se deberán activar las ponderaciones, indicando la variable que contiene las frecuencias o ponderaciones*.

EJEMPLO

*Con los datos del **anexo 3**, que hacen referencia a las llegadas de vuelos al aeropuerto de Barajas en un día determinado, obtener los siguientes estadísticos: media y desviación típica de las variables Retraso y Pasaje ponderadas por la variable Núm.*

En primer lugar, se activa la ponderación adecuada. Con la secuencia *Datos* > *Ponderar casos* se indica, en el cuadro de diálogo que aparece, la variable que recoge las frecuencias: Num.

Con la secuencia *Analizar* > *Estadísticos Descriptivos* > *Descriptivos* se abre un cuadro de diálogo donde se seleccionan las variables Retraso y Pasaje y en el cuadro *Opciones* se activan los estadísticos correspondientes.

Estadísticos descriptivos

	N	Mínimo	Máximo	Suma	Media	Desv. típ.
PASAJE Núm medio pasajeros	344	89	205	49295	143,30	30,00
RETRASO Retraso en mn.	344	0	60	5180	15,06	17,51
N válido (según lista)	344					

En base a estos resultados se observa que la media de pasajeros por vuelo es 143,3 y que el retraso medio es de 15 minutos. La desviación típica del Pasaje es 30 que representa una dispersión baja con respecto a su media. En cambio, la desviación del retraso es 17,51 y representa un grado de dispersión elevado con respecto a la media que es 15,06 minutos. En consecuencia, la media del Pasaje es más representativa que la del Retraso.

PREGUNTAS

- **¿Cómo se realiza la distribución de frecuencias?**
- **¿Qué gráficos se realizan para la estadística descriptiva?**

Tercera unidad

ANÁLISIS DE DATOS

3.1 ANÁLISIS EXPLORATORIO DE DATOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo realizará adecuadamente un análisis exploratorio de datos.

ANÁLISIS EXPLORATORIO DE DATOS

El análisis exploratorio tiene como objetivo identificar el modelo teórico más adecuado para representar la población de la cual proceden los datos muestrales. Dicho análisis se basa en gráficos y estadísticos que permiten explorar la distribución identificando características tales como: valores atípicos o outliers, saltos o discontinuidades, concentraciones de valores, forma de la distribución, etc. Por otra parte, este análisis se puede realizar sobre todos los casos conjuntamente o de forma separada por grupos. En este último caso los gráficos y estadísticos permiten identificar si los datos proceden de una o varias poblaciones, considerando la variable que determina los grupos como factor diferenciador de las poblaciones. También permite comprobar, mediante técnicas gráficas y contrastes no paramétricos, si los datos han sido extraídos de una población con distribución aproximadamente normal.

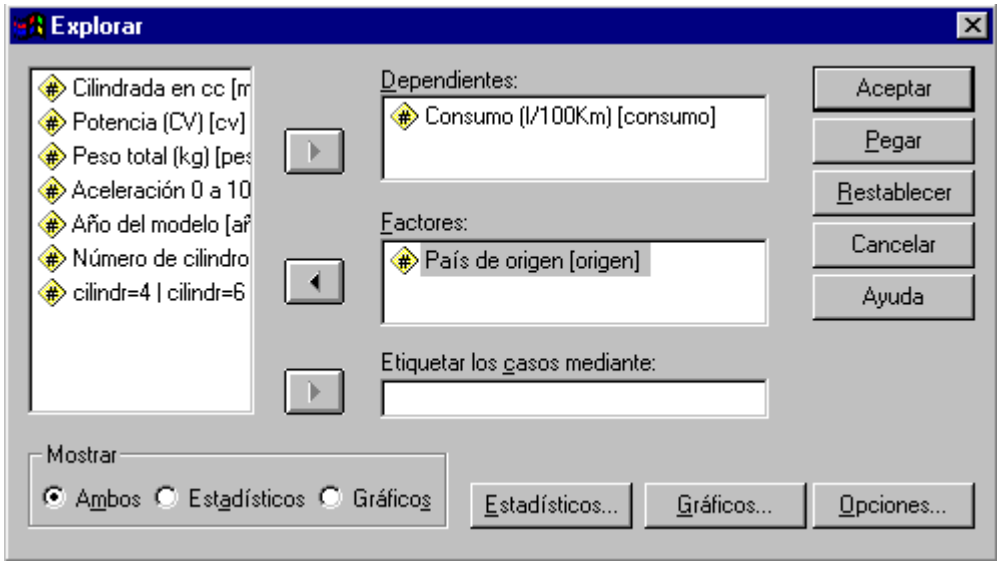
Para realizar un análisis exploratorio, la secuencia de instrucciones es:

Analizar

Estadísticos

Descriptivos

Explorar

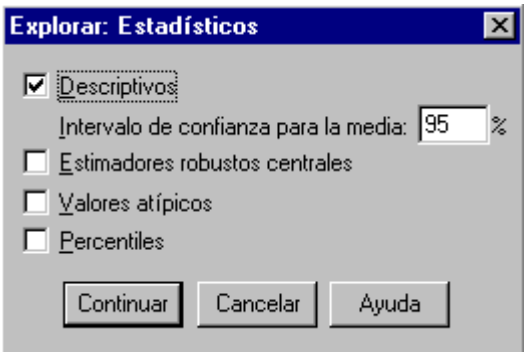


• Si el análisis de la variable se realiza conjuntamente para todos los casos es suficiente indicar la o las variables en la ventana *Dependientes*.

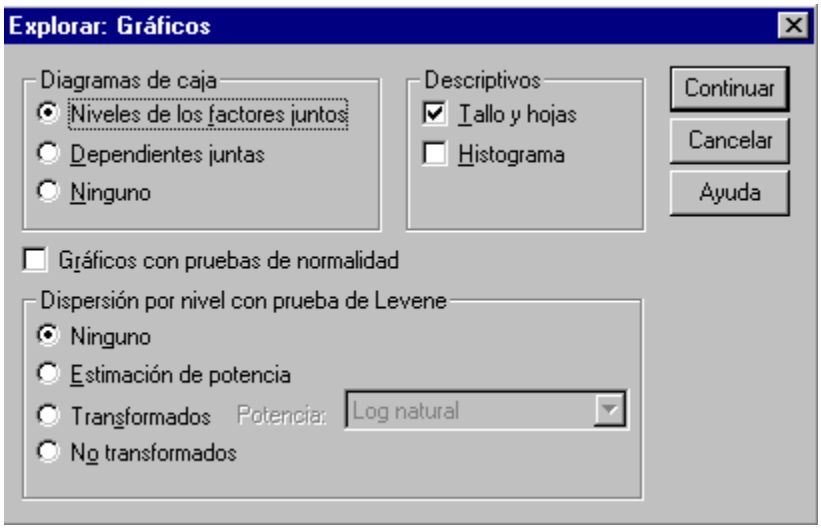
• Si el análisis de la variable se realiza por grupos es necesario indicar también la variable que define los grupos en la ventana *Factores*.

Opcionalmente se puede indicar en la ventana *Etiquetar los casos mediante* una variable cuyos valores se tomarán para etiquetar los outliers.

El análisis exploratorio calcula, por defecto, los estadísticos más importantes así como el intervalo de confianza para la media al 95%, el gráfico de tallo y hojas y el diagrama de caja. Para ampliar éste análisis se puede acceder a los siguientes cuadros de diálogo intervalo media:



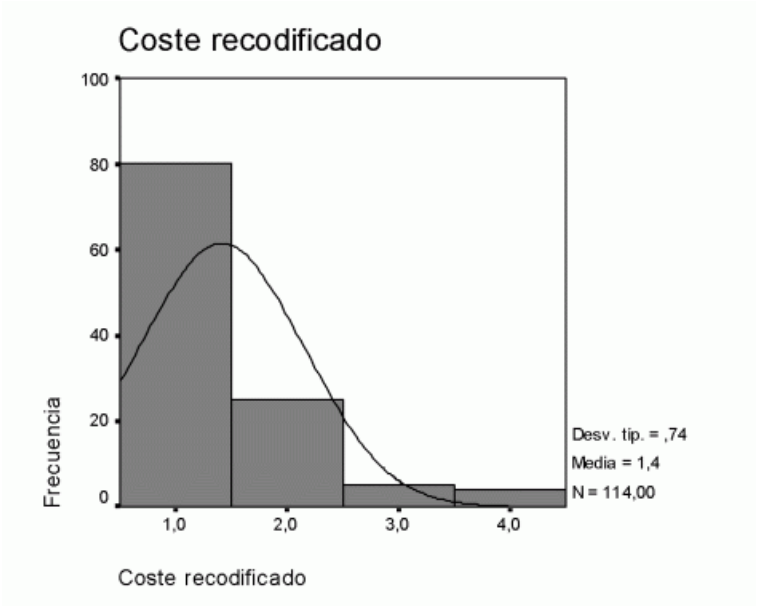
• *Estadísticos*: Permite modificar el grado de confianza del intervalo para la media, calcular *Estimadores robustos centrales* (estimador M de Huber, estimador en onda de Andrews, estimador M redescendente de Hampel, estimador biponderado de Tukey), y hallar los *Valores atípicos* (se obtienen los 5 mayores y los 5 menores valores de la distribución) y algunos *Percentiles* (los cuartiles y el 5º, 10º, 90º y 95º centil).



• *Gráficos*:

- Las opciones del *Diagrama de caja* se utilizan sólo cuando se han seleccionado varias variables dependientes. Por defecto, se presentan en gráficos distintos las variables dependientes seleccionadas, y para cada una de ellas, en el mimo gráfico, las cajas de los distintos grupos definidos por el factor. Si se selecciona *Dependientes juntas* se representan en un único gráfico las cajas correspondientes a todas las variables dependientes. Con la opción *Ninguno* se omite la presentación de los diagramas de caja.
- Las alternativas de *Descriptivos* son el gráfico de tallo y hojas, activado por defecto, y el histograma. Estos gráficos se elaboran por separado para todos los grupos definidos para cada una de las variables dependientes.
- Si se activa la opción *Gráficos con pruebas de normalidad* se obtienen para cada una de las variables dependientes y para cada uno de los grupos el correspondiente gráfico Q-Q Normal y el gráfico Q-Q Normal sin tendencia. Estos gráficos permiten comprobar si las poblaciones de las que se han extraído las muestras presentan distribución normal. El Q-Q Normal presenta simultáneamente para cada elemento el valor observado y el valor esperado bajo el supuesto de normalidad. Si los datos proceden de una distribución normal los puntos aparecen agrupados en torno a la línea recta esperada. El Q-Q Normal sin tendencia se basa en las diferencias entre los valores observados y los valores esperados bajo la hipótesis de normalidad. Si estas diferencias se distribuyen aleatoriamente alrededor del eje de abscisas puede suponerse que la hipótesis de normalidad es sostenible. Además, esta opción permite contrastar la hipótesis de normalidad con las prueba de Kolgomorov-Smirnov* y de Shapiro-Wilks*.
- La opción *Dispersión por nivel con prueba de Levene*, activando *No transformados*, permite contrastar la hipótesis de igualdad de varianza para los grupos definidos por un factor.

• *Opciones* controla el tratamiento de los valores missing en el análisis exploratorio.



- Por defecto está activada la opción Excluir casos según lista con la que se eliminan de todos los cálculos y gráficos los casos que presentan algún valor missing, ya sea en alguna de las variables dependientes o en algún factor. Con esta opción el número de casos válidos es el mismo en todos los resultados.

- Si se activa Excluir casos según pareja, los casos que presentan algún valor missing en alguna variable dependiente sólo se excluyen en los cálculos de estadísticos correspondientes a dicha variable, y sí que se incluyen en los de otras variables dependientes. Con esta opción el número de casos válidos no tiene porque coincidir en todos los resultados.

- Con la opción Mostrar valores, los valores missing de un factor definen un nuevo grupo de casos. Si los valores missing son de la variable dependiente éstos no se incluyen en el cálculo de los estadísticos.

El análisis exploratorio de datos puede aplicarse a variables cuantitativas. La variable factor debe presentar un número limitado de categorías y es conveniente expresarlas numéricamente o con una cadena alfanumérica corta.

EJEMPLOS

Ejemplo 1. *Con la base de datos **Enctran.sav** realizar el análisis exploratorio básico de las variables: Alt y Peso, diferenciando por las variables Genero y Curso. Etiquetar los valores atípicos con la variable Num.*

En el cuadro de diálogo *Explorar* se seleccionan las siguientes variables:

Con ello se obtendrá el análisis exploratorio básico de las variables dependientes (Altura y Peso). Los factores Género y Curso son variable cualitativas con dos modalidades cada una; por lo tanto, para cada dependiente quedarán definidos dos grupos con respecto al género (Hombres y Mujeres) y dos con respecto al curso (Primero y Segundo). La variable Num se selecciona para etiquetar los valores outliers.

Los resultados que se obtienen con las opciones activadas por defecto son:

• Descriptivos:

Contiene los valores de los estadísticos más utilizados para las variables Altura y Peso en función de los grupos inducidos por las variables factores. En la tabla siguiente se recogen los resultados de ambas variables correspondientes a los grupos asociados al factor Genero. El lector puede comprobar que el output contiene también el cuadro análogo correspondiente a los grupos definidos por elfactor Curso.

Descriptivos

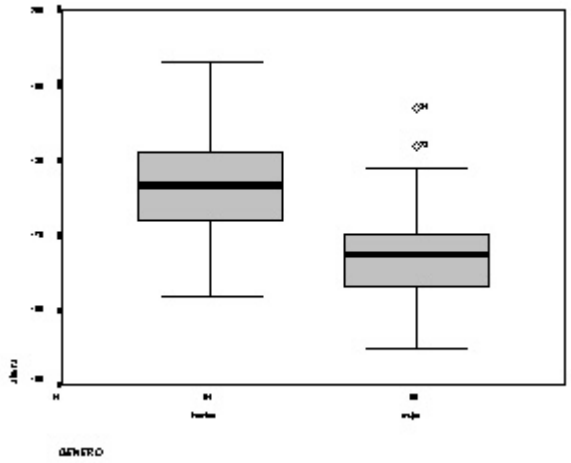
	GENERO			Estadístico	Error típ.
altura	hombre	Media		176,94	,92
		Intervalo de confianza para la media al 95%	Límite inferior	175,11	
			Límite superior	178,78	
		Media recortada al 5%		176,88	
		Mediana		176,50	
		Varianza		45,223	
		Desv. típ.		6,72	
		Mínimo		162	
		Máximo		193	
		Rango		31	
		Amplitud intercuartil		9,25	
		Asimetría		,217	,325
		Curtosis		,096	,639
	mujer	Media		167,50	,81
		Intervalo de confianza para la media al 95%	Límite inferior	165,89	
			Límite superior	169,11	
		Media recortada al 5%		167,24	
		Mediana		167,50	
		Varianza		39,000	
		Desv. típ.		6,24	
		Mínimo		155	
		Máximo		187	
		Rango		32	
		Amplitud intercuartil		7,00	
		Asimetría		,739	,309
		Curtosis		,886	,608
peso	hombre	Media		70,52	1,34
		Intervalo de confianza para la media al 95%	Límite inferior	67,82	
			Límite superior	73,22	
		Media recortada al 5%		70,42	
		Mediana		70,00	
		Varianza		97,613	
		Desv. típ.		9,88	
		Mínimo		45	
		Máximo		95	
		Rango		50	
		Amplitud intercuartil		10,25	
		Asimetría		,103	,325
		Curtosis		,483	,639
	mujer	Media		58,93	1,13
		Intervalo de confianza para la media al 95%	Límite inferior	56,68	
			Límite superior	61,19	
		Media recortada al 5%		58,61	
		Mediana		58,00	
		Varianza		76,368	
		Desv. típ.		8,74	
		Mínimo		40	
		Máximo		81	
		Rango		41	
		Amplitud intercuartil		12,00	
		Asimetría		,566	,309
		Curtosis		,073	,608

• Gráficos:

En el visor de resultados se obtienen los gráficos de tallo y hoja y los diagramas de caja.

```
altura Stem-and-Leaf Plot for
GENERO= hombre

 Frequency    Stem &  Leaf

     2.00       16 .  23
     2.00       16 .  59
    15.00       17 .  000000122223444
    15.00       17 .  555555567788888
    12.00       18 .  000000122233
     5.00       18 .  55568
     3.00       19 .  023

 Stem width:   10
 Each leaf:       1 case(s)


altura Stem-and-Leaf Plot for
GENERO= mujer

 Frequency    Stem &  Leaf

      .00       15 .
     4.00       15 .  5799
    16.00       16 .  0000122233334444
    20.00       16 .  55556667778888888999
    12.00       17 .  000000000134
     6.00       17 .  588889
     2.00 Extremes    (>=182)

 Stem width:   10
 Each leaf:       1 case(s)
```

Como puede observarse la distribución de la variable Altura para el grupo mujer presenta dos outliers o valores extremos, con valores superiores a 182 cm. Gráficos similares se obtienen también para el resto de combinaciones variable-factor.



En el diagrama de caja anterior se observa que el valor central de la distribución de la variable Altura es notablemente superior en el grupo de hombres; la distribución de la altura en ambos grupos es prácticamente simétrica y, aparentemente, presentan dispersiones parecidas. Obsérvese que los valores outliers están etiquetados con el número de caso.

Ejemplo *Para las mismas variables del ejemplo anterior y diferenciando únicamente por el factor género comprobar:*

2.

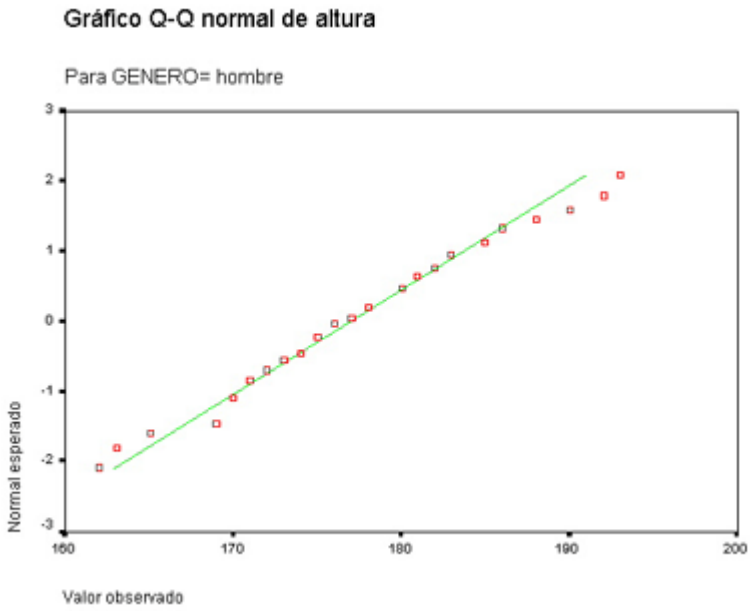
• La hipótesis de que las muestras provienen de poblaciones normales; En el cuadro de diálogo que se abre con la secuencia *Analizar* > *Explorar* > *Gráficos* se activa la opción *Gráficos con pruebas de normalidad*:

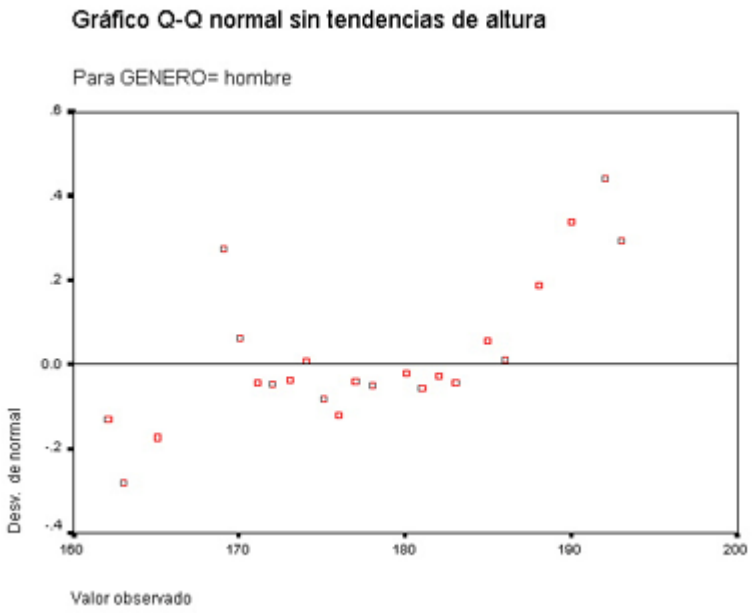
Pruebas de normalidad

		Kolmogorov-Smirnov[a]		
	GENERO	Estadístico	gl	Sig.
altura	hombre	,095	54	,200*
	mujer	,161	60	,001
peso	hombre	,103	54	,200*
	mujer	,151	60	,002

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors





El estadístico del contraste Kolmogorov-Smirnov para la variable Altura en el grupo hombres toma el valor 0,95 que no permite rechazar la hipótesis nula de normalidad para niveles de significación inferiores a 0,2. En el grupo de mujeres, por el contrario, el estadístico toma el valor 0,161 con el que se rechaza la hipótesis de normalidad para niveles de significación superiores a 0,001. De la misma forma se interpretan los resultados correspondientes a la variable peso.

En el gráfico Q-Q normal de la variable Altura se observa, para el grupo de los hombres, que los puntos están situados casi sobre la línea recta lo cual es un indicio

de normalidad de la población de origen. Este resultado es compatible con el del contraste de Kolmogorov-Smirnov.

• La hipótesis de que las muestras provienen de poblaciones con igual varianza.

Esta prueba debe realizarse cuando se desea contrastar hipótesis referidas a las medias de dos o más poblaciones definidas mediante una variable factor.

En el cuadro de diálogo *Explorar: Gráficos* se activa la opción *No transformados* del recuadro *Dispersión por nivel con prueba de Levene*.

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
altura	Basándose en la media	,651	1	112	,421
	Basándose en la mediana.	,648	1	112	,423
	Basándose en la mediana y con gl corregido	,648	1	111,999	,423
	Basándose en la media recortada	,650	1	112	,422
peso	Basándose en la media	,214	1	112	,644
	Basándose en la mediana.	,243	1	112	,623
	Basándose en la mediana y con gl corregido	,243	1	108,912	,623
	Basándose en la media recortada	,228	1	112	,634

El estadístico de Levene, en todos los casos, permite no rechazar la hipótesis de homogeneidad de la varianza (obsérvese que los niveles de significación para los que se rechazaría esta hipótesis son todos superiores a 0,4, por lo tanto, para los niveles de significación habituales no se rechaza la hipótesis nula).

PREGUNTAS

- **¿Cómo se realiza las tablas de contingencia?**
- **¿Para que tipo de variables se emplean los Box-Plot?**
- **Realice un Histograma de Frecuencias y un gráfico de dispersión.**

3.2 ANÁLISIS DE DATOS EXPLICATIVOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo realizará adecuadamente un análisis de datos explicativos, asociación de datos nominales y ordinales.

TABLAS DE CONTINGENCIA

Los datos procedentes de la observación de dos variables categóricas o categorizadas se ordenan en una tabla de contingencia. ésta consiste en una tabla de doble entrada con **I** filas y **J** columnas,siendo **I** y **J** el número de categorías de cada una de las variables. La casilla o celda, situada en la intersección de una fila y una columna dada, recoge la frecuencia absoluta que presentan simultáneamente las modalidades que ocupan las correspondientes fila y columna.

Para obtener una tabla de contingencia la secuencia es la siguiente:

Analizar

Estadísticos Descriptivos

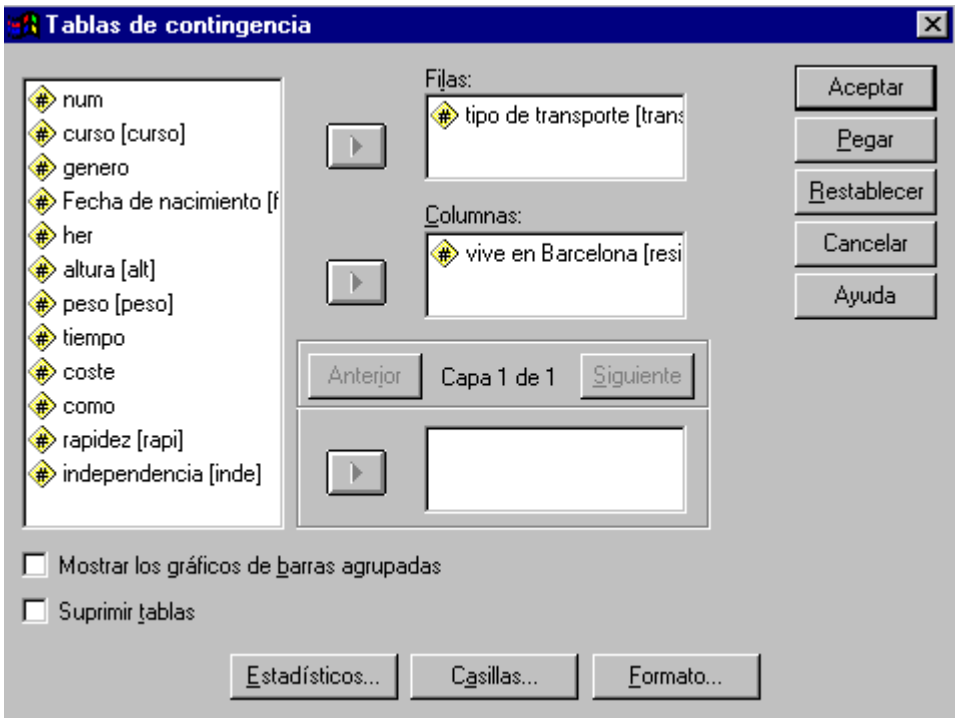
Tablas de contingencia

En el cuadro de diálogo se seleccionan las variables cuyo análisis conjunto se desea realizar.

Si se acepta sin modificar ninguna opción se obtiene la tabla de contingencia con las frecuencias conjuntasabsolutas y las marginales.

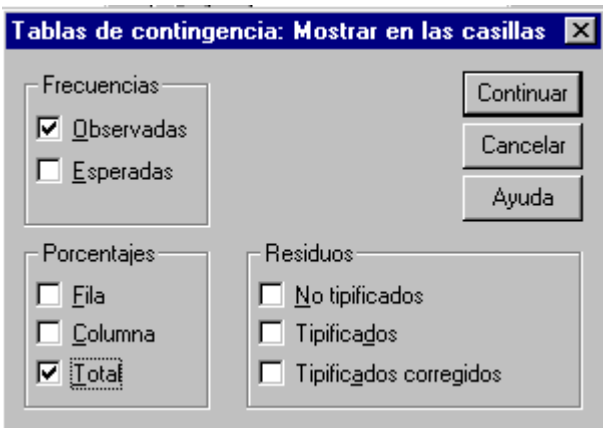
Si se activa *Mostrar los gráficos de barras agrupadas* se obtiene un diagrama en donde para cada una de las modalidades de la variable fila se representa gráficamente, mediante una barra, la frecuencia conjunta con cada una de las modalidades de la variable columna.

Si sólo se quiere obtener algún estadístico sin necesidad de visualizar la tabla de contingencia se debe activar la opción *Suprimir tablas*.



FRECUENCIAS RELATIVAS Y CONDICIONADAS

Para incluir en la tabla de contingencia las frecuencias relativas o las frecuencias condicionadas se debe activar el botón *Casillas* que abre el siguiente cuadro de diálogo:



En este cuadro está activada por defecto la opción *Frecuencias Observadas* que como ya se ha dicho proporciona la frecuencia absoluta conjunta, es decir, el número de casos que presentan simultáneamente las modalidades fila y columna. Si se activa *Porcentajes Total* en cada casilla de la tabla aparece la frecuencia relativa conjunta en porcentaje o proporción de casos que presentan simultáneamente las modalidades fila y columna sobre el total observado.

Si se activa *Porcentajes Fila y/o Columna*, la tabla de contingencia recoge en cada casilla, junto a la frecuencia conjunta absoluta y relativa, los porcentajes condicionados al total de la fila y/o al total de la columna.

FRECUENCIAS ESPERADAS

La opción *Casillas* también permite obtener simultáneamente las frecuencias absolutas observadas y las esperadas bajo el supuesto de independencia de las variables. Dos variables *X* e *Y* son estadísticamente independientes cuando la frecuencia relativa conjunta *(fij)* coincide con el producto de las frecuencias relativas marginales (frecuencias de sus distribuciones unidimensionales fi. fj) para todos los valores de *X* e *Y*.

X e Y son estadísticamente independientes $\Longleftarrow f_{ij} = f_{i.} f_{j} \quad \forall i, j$

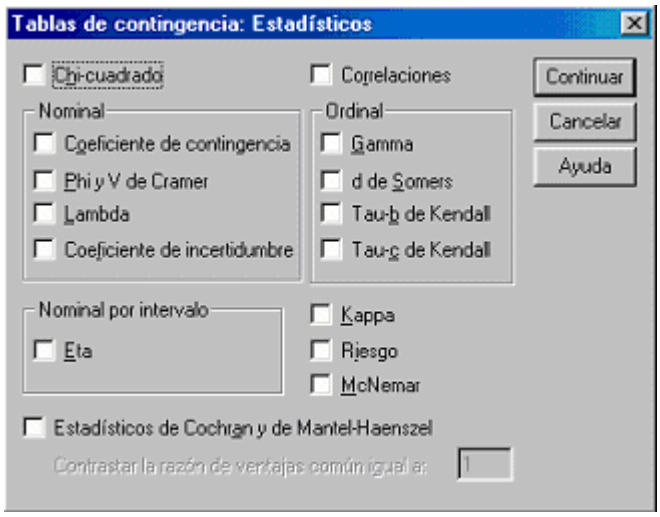
La condición de independencia implica que las variables no se condicionan y, por lo tanto, las frecuencias condicionadas (fila o columna) coinciden con las marginales (de X o de Y) en términos relativos. Si se supone independencia entre dos variables se espera que las frecuencias relativas conjuntas sean iguales al producto de las marginales y sus frecuencias absolutas esperadas serán:

$$e_{ij} = \frac{n_{i.} n_{.j}}{n} \quad \forall i, j$$

Para incluir esta frecuencia en la tabla de contingencia se activa la opción *Frecuencias Esperadas*.También se pueden obtener los resíduos o diferencias entre las frecuencias observadas y las esperadas, activando *Resíduos No tipificados*. Estos resíduos evalúan las discrepancias entre lo observado y lo que se espera cuando las variables son independientes, y a partir de ellos se calculan la mayoría de medidas de asociación de variables cualitativas.

MEDIDAS DE ASOCIACIÓN PARA DATOS NOMINALES

El botón *Estadísticos* permite calcular las medidas de asociación más utilizadas para variables nominales y ordinales. Por defecto no hay ninguna opción activada.



Algunos de los estadísticos que recoge este cuadro de diálogo son:

1. Chi-cuadrado, con esta opción se obtienen los estadísticos:

- Chi-cuadrado de Pearson: $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$ (donde *nij* y *eij* son las frecuencias absoluta observada y esperada, respectivamente). Si la tabla es 2x2, el estadístico se puede calcular mediante la expresión abreviada $\chi^2 = \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$ (donde *ni.* y *nj.* son las frecuencias marginales por filas y columnas). Este estadístico es fiable únicamente cuando por lo menos el 80% de las frecuencias esperadas son mayores que 5.
- Corrección de continuidad (de Yates):en las tablas 2x2 corrige el error que se comete al calcular el estadístico Chi-cuadrado de Pearson debido a la aproximación de una distribución discreta por una continua. El estadístico corregido es $\chi^2 = \frac{n(|n_{11}n_{22}-n_{12}n_{21}|-0,5n)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$.
- Contraste de la razón de verosimilitud:se basa en el estadístico G^2 que se calcula como $G^2 = -2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\ln\left(\frac{e_{ij}}{n_{ij}}\right)$.
- Prueba exacta de Fisher:si la tabla es 2x2 y los totales marginales se consideran fijos, la probabilidad exacta de obtener la tabla analizada bajo el supuesto de independencia es:

$$p = \frac{(n_{11}+n_{12})!(n_{21}+n_{22})!(n_{11}+n_{21})!(n_{12}+n_{22})!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}.$$

Si la variable es *Nominal* los estadísticos que se pueden calcular son los siguientes:

2. Coeficiente de Contingencia de Pearson: se define como $C=\sqrt{\frac{\chi^2}{\chi^2+n}}$. Esta medida de asociación no se ve afectada por el tamaño de la muestra y está acotada, $0 \leq C < 1$. Si las variables son independientes se tiene C=0, pero en caso de asociación perfecta nunca alcanza el valor 1. Para tablas cuadradas RxR se puede calcular la cota superior que es $\sqrt{\frac{R-1}{R}}$; en tal caso $0 \leq C \leq \sqrt{\frac{R-1}{R}}$.

3. Phi y V de Cramer:

- El coeficiente Phi es $\Phi = \sqrt{\frac{\chi^2}{n}}$. Este coeficiente toma el valor 0 en caso de independencia, pero su valor máximo depende de las dimensiones de la tabla; para tablas superiores a 2xJ o Ix2 el coeficiente Φ puede tomar valores mayores que 1.

- El coeficiente V de Cramer se calcula también en función de χ^2 :

$V=\sqrt{\frac{\chi^2}{n(k-1)}}$ (donde k =mín(I,J)) y siempre está acotado entre 0 y 1 (sea cual sea la dimensión de la tabla). Para tablas 2xJ o Ix2 el valor de V coincide con el de Φ.

4. *Lambda:* Incluye la λ simétrica y asimétrica y la τ de Goodman y Kruskal. Ambas medidas se basan en la reducción proporcional del error de predicción cuando se utilizan los valores de la variable independiente para predecir los de la dependiente.

- El coeficiente λ de Kruskal se calcula:

- Si la X_1 es la variable dependiente:

$$\lambda_{x1} = \frac{(f_{m1} + f_{m2} + \ldots + f_{mJ}) - f_{m.}}{1 - f_{m.}}$$

donde *fmj* es la frecuencia relativa conjunta máxima correspondiente a la columna j-ésima y $f_{m.}$. es la frecuencia relativa marginal máxima de X_1. El valor de este estadístico, acotado entre 0 y 1, indica en cuanto se reduce la incertidumbre de la variable X_1 cuando se dispone de información sobre el comportamiento de X_2.

- De forma análoga se calcula λ_{x2}.

- Si no se puede determinar cual de las dos variables es la dependiente y cual la independiente se calcula el coeficiente λ simétrica como:

$$\lambda = \frac{(1 - f_{m.})\lambda_{x1} + (1 - f_{.m})\lambda_{x2}}{2 - (f_{m.} + f_{.m})}$$

El valor de λ simétrica está comprendido entre λ_{x1} y λ_{x2}.

Un valor λ igual a 0 indica que la información acerca de la variable independiente no ayuda en absoluto a predecir los valores de la variable dependiente; mientras que un valor λ igual a 1 indica asociación predictiva perfecta entre las variables.

- La τ de Goodman y Kruskal indica la reducción en el error de clasificación de los elementos para una de las variables (dependiente) cuando se tiene información sobre el comportamiento de la otra (independiente). Si X_1 es la variable dependiente, se calcula: $\tau_1 = \frac{TMCSI - TMCCI}{TMCSI}$ siendo TMCSI=total de elementos mal clasificados sin información acerca de X_2. y TMCCI=total de elementos mal clasificados con información acerca de X_2. La τ_2 se define de manera análoga.

5. *Coeficiente de incertidumbre, U*: mide la reducción proporcional del error cuando los valores de una variable se emplean para predecir valores de la otra.

En el cuadro *Nominal por intervalo* se encuentra el estadístico:

6. *Eta:* Este coficiente es apropiado cuando la variable dependiente se mide en una escala de intervalo, mientras que la independiente es una variable nominal. El programa muestra dos valores de *Eta* según se considere que la variable medida en una escala de intervalo esté en las filas o en las columnas.

MEDIDAS DE ASOCIACIÓN PARA DATOS ORDINALES

En el cuadro de diálogo *Tablas de contingencia*: *Estadísticos* pueden activarse diversas opciones que proporcionan medidas de asociación cuando las variables se miden por lo menos en una escala ordinal; las más utilizadas son:

1. *Correlaciones*: con esta opción se obtienen los estadísticos:

- *Coeficiente de correlación de Pearson*: es una medida de asociación lineal adecuada para variables medidas en escala de intervalo *.
- *Coeficiente de correlación de Spearman*: mide el grado de correspondencia que existe entre los rangos que se asignan a los valores de las variables analizadas. Por ello, este coeficiente se puede calcular con datos ordinales, y se define: $r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, siendo *d*i la diferencia entre los rangos correspondientes a la observación i-ésima. El coeficiente toma valores entre -1 y +1. Un valor cercano a 0 indica que las variables apenas están relacionadas.

El cuadro *Ordinal* recoge una serie de estadísticos basados en el número de concordancias y discordancias que aparecen al comparar las puntuaciones asignadas a los mismos casos según dos criterios (o jueces) diferentes. Así, por ejemplo, si X_1 recoge las puntuaciones asignadas a los casos según el primer criterio y X_2, según el segundo, para la obtención de concordancias y discordancias que aparecen entre los dos criterios, se procede de la siguiente forma:

- se ordenan los pares de puntuaciones de acuerdo con el orden natural de las puntuaciones asignadas según el primer criterio, X_1.
- se compara cada valor de X_2, con cada uno de los que le siguen, y se registra una concordancia (+1) cuando los dos valores siguen el orden natural, una discordancia (-1) cuando el orden está invertido y un empate (0) cuando coinciden ambas puntuaciones.
- se calculan C total de las concordancias, D total de las discordancias y E el número total de empates.

El número total de comparaciones es $\frac{n(n-1)}{2}$ incluyendo empates.

1. *Gamma*: El estadístico Gamma se define como $\frac{C-D}{C+D}$. Este análisis excluye los casos que presentan la misma puntuación en las dos variables (empates).

2. *Tau-b de Kendall*. Este coeficiente incluye los empates contemplando por separado los que aparecen en la variable $X_1\ (E_{x1})$ y los que aparecen en la variable $X_2\ (E_{x2})$.

Se define como $\tau_b = \frac{C-D}{\sqrt{(C+D+E_{x1})(C+D+E_{x2})}}$.

3. *Tau-c de Kendall*. Este estadístico se define como $\tau_c = \frac{2\,k\,(C-D)}{n^2(k-1)}$ siendo k el menor número de casos no empatados que presentan X_1 o X_2.

4. *d de Somers*: A diferencia de los anteriores este estadístico considera que las variables pueden ser simétricas o dependientes. En el primer caso, el estadístico *d de Somers* coincide con la *Tau-b de Kendall*. En el segundo supuesto, se diferencia del estadístico *Gamma* en que incluye los empates de la variable que considera dependiente. Si la variable dependiente es X_1, $d = \frac{C-D}{C+D+E_{x1}}$.

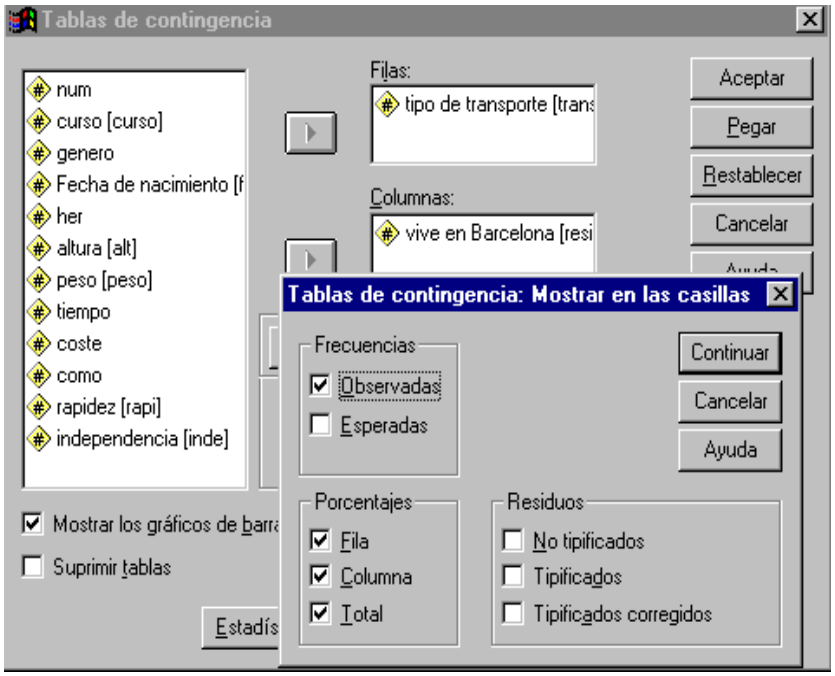
Todas estas medidas toman valores entre -1 y +1, y alcanza los valores extremos cuando existe concordancia o discordancia perfecta. Valores próximos a 0 indican ausencia de asociación.

EJEMPLO

Ejemplo 1.

Obtenga una tabla de contingencia con las frecuencias absolutas, relativas y condicionadas para las variables y el correspondiente diagrama de barras.

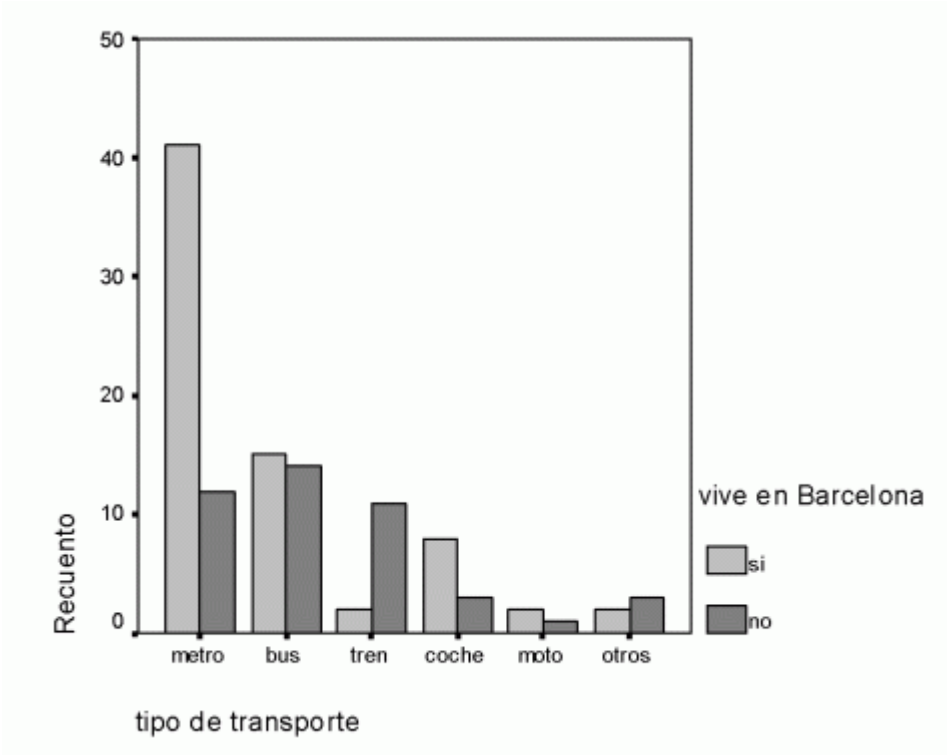
En el cuadro de diálogo *Tablas de contingencia* se activa la opción *Mostrar los gráficos de barras agrupadas*, y en el cuadro de diálogo *Tablas de contingencia* > *Casillas* se activan las opciones *Porcentajes Fila, Columna y Total*.



Los resultados que se obtienen son los siguientes:

Tabla de contingencia tipo de transporte * vive en Barcelona

			vive en Barcelona		
			si	no	Total
tipo de transporte	metro	Recuento	41	12	53
		% de tipo de transporte	77,4%	22,6%	100,0%
		% de vive en Barcelona	58,6%	27,3%	46,5%
		% del total	36,0%	10,5%	46,5%
	bus	Recuento	15	14	29
		% de tipo de transporte	51,7%	48,3%	100,0%
		% de vive en Barcelona	21,4%	31,8%	25,4%
		% del total	13,2%	12,3%	25,4%
	tren	Recuento	2	11	13
		% de tipo de transporte	15,4%	84,6%	100,0%
		% de vive en Barcelona	2,9%	25,0%	11,4%
		% del total	1,8%	9,6%	11,4%
	coche	Recuento	8	3	11
		% de tipo de transporte	72,7%	27,3%	100,0%
		% de vive en Barcelona	11,4%	6,8%	9,6%
		% del total	7,0%	2,6%	9,6%
	moto	Recuento	2	1	3
		% de tipo de transporte	66,7%	33,3%	100,0%
		% de vive en Barcelona	2,9%	2,3%	2,6%
		% del total	1,8%	,9%	2,6%
	otros	Recuento	2	3	5
		% de tipo de transporte	40,0%	60,0%	100,0%
		% de vive en Barcelona	2,9%	6,8%	4,4%
		% del total	1,8%	2,6%	4,4%
Total		Recuento	70	44	114
		% de tipo de transporte	61,4%	38,6%	100,0%
		% de vive en Barcelona	100,0%	100,0%	100,0%
		% del total	61,4%	38,6%	100,0%



En la tabla de contingencia se observan, entre otros resultados, los siguientes:

- Un total de 53 personas utilizan el metro de las cuales 41 viven en Barcelona y 12 no.

- El 36% del total de casos de la muestra utilizan el metro y viven en Barcelona.

- El 58,6% de los que viven en Barcelona utilizan el metro.

- El 77,4% de los que utilizan el metro viven en Barcelona.

- El 10,5% del total de casos utilizan el metro y no viven en Barcelona.

- El 27,3% de los que no viven en Barcelona utilizan el metro.

-El 22,6% de los que utilizan el metro no viven en Barcelona.

Análogamente se interpretan el resto de resultados.

Ejemplo 2.

Obtenga la tabla de contingencia para las variables del ejemplo anterior con frecuencias observadas, esperadas y residuos no tipificados.

Tabla de contingencia tipo de transporte * vive en Barcelona

			vive en Barcelona		
			si	no	Total
tipo de transporte	metro	Recuento	41	12	53
		Frecuencia esperada	32,5	20,5	53,0
		Residual	8,5	-8,5	
	bus	Recuento	15	14	29
		Frecuencia esperada	17,8	11,2	29,0
		Residual	-2,8	2,8	
	tren	Recuento	2	11	13
		Frecuencia esperada	8,0	5,0	13,0
		Residual	-6,0	6,0	
	coche	Recuento	8	3	11
		Frecuencia esperada	6,8	4,2	11,0
		Residual	1,2	-1,2	
	moto	Recuento	2	1	3
		Frecuencia esperada	1,8	1,2	3,0
		Residual	,2	-,2	
	otros	Recuento	2	3	5
		Frecuencia esperada	3,1	1,9	5,0
		Residual	-1,1	1,1	
Total		Recuento	70	44	114
		Frecuencia esperada	70,0	44,0	114,0

La observación de los resíduos permite tener una primera aproximación sobre la existencia de asociación entre las variables.Si los resíduos en valor absoluto son próximos a 0 se espera que la hipótesis de independencia entre las variables no se pueda rechazar. Por el contrario, cuanto mayores sean los valores absolutos de los resíduos se tendrán más indicios sobre la existencia de asociación. En cualquier caso, la confirmación de la existencia o no de asociación entre las variables se obtiene a partir de los estadísticos correspondientes. En este ejemplo se observan unos valores de los resíduos, en general, elevados, lo que nos lleva a pensar en la existencia de asociación entre el tipo de transporte y el lugar de residencia.

Ejemplo 3.

Otenga el coeficiente Chi-cuadrado, coeficiente de contingencia, Phi y V de Cramer e interprete los resultados.

En el cuadro de diálogo *Tablas de contingencia*: *Estadísticos* se activan las opciones correspondientes. Los resultados aparecen en los siguientes cuadros.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	20,052[a]	5	,001
Razón de verosimilitud	20,584	5	,001
Asociación lineal por lineal	4,558	1	,033
N de casos válidos	114		

a. 5 casillas (41,7%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 1,16.

Como se ha visto en el ejemplo anterior, la tabla de contingencia de las variables Trans y Resi presenta 5 casillas con frecuencia esperada inferior a 5, lo que representa un 41,7% de las casillas. En estas condiciones, los resultados del contraste Chi-cuadrado no son fiables. Para resolver este problema se agrupan clases hasta obtener frecuencias esperadas superiores a 5. Agrupando Moto y Otros se obtiene una clase que tiene frecuencia esperada prácticamente igual a 5. Para agruparlas se recodifica la variable Trans en una nueva variable (Trans2) manteniendo todos los valores de Trans excepto los correspondientes a Moto y Otros, a los que se les asigna un único valor.

Repitiendo el análisis se obtiene:

Tabla de contingencia TRANS2 * vive en Barcelona

			vive en Barcelona		
			si	no	Total
TRANS2	Metro	Recuento	41	12	53
		Frecuencia esperada	32,5	20,5	53,0
	Bus	Recuento	15	14	29
		Frecuencia esperada	17,8	11,2	29,0
	Tren	Recuento	2	11	13
		Frecuencia esperada	8,0	5,0	13,0
	Coche	Recuento	8	3	11
		Frecuencia esperada	6,8	4,2	11,0
	Otros	Recuento	4	4	8
		Frecuencia esperada	4,9	3,1	8,0
Total		Recuento	70	44	114
		Frecuencia esperada	70,0	44,0	114,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	19,490[a]	4	,001
Razón de verosimilitud	20,043	4	,000
Asociación lineal por lineal	3,728	1	,054
N de casos válidos	114		

a. 3 casillas (30,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,09.

Como puede observarse en la tabla de contingencia hay tres frecuencias esperadas menores que 5, es decir, el 30% del total; no obstante una de dichas frecuencias es igual a 4,9 que a efectos prácticos puede considerarse igual a 5. Por tanto sólo el 20% de las frecuencias esperadas es inferior a 5 y en estas condiciones los resultados del contraste Chi-cuadrado son fiables. El valor del estadístico Chi-cuadrado* es 19,490 y la razón de verosimilitud 20,043. Estos valores difieren significativamente de 0 para niveles de significación superiores a 0,001, lo que significa que se rechaza la hipótesis de independencia para los niveles de significación habituales.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,413	,001
	V de Cramer	,413	,001
	Coeficiente de contingencia	,382	,001
N de casos válidos		114	

a. No asumiendo la hipótesis nula.

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

***** NOTA A PIE ** Obsérvese que no se ha realizado la corrección de continuidad de Yates, porque la tabla no es de dimensiones 2x2.***

-Las medidas basadas en el estadístico Chi-cuadrado, que son los coeficientes Phi y V de Cramer, toman ***ambas**** el valor 0,413 con un nivel de significación crítico 0,001; lo que implica la existencia de asociación moderadamente fuerte entre las variables.

** ***NOTA A PIE ** En este caso, ambos coeficientes coinciden por tratarse de una tabla de dimensiones Ix2.***

-El coeficiente de contingencia de Pearson toma el valor 0,382, también con el nivel de significación 0,001. Esto es consistente con el valor de los coeficientes anteriores.

Ejemplo 4:

Analizar si existe asociación entre las variables Tipo e Internet del archivo ***Encinf.sav*** *e indicar si es posible establecer una relación de dependencia entre ambas.*

Para analizar si existe asociación e indicar el tipo de relación de dependencia se activan las opciones Chi-cuadrado, *Coeficiente de contingencia, Phi y V de Cramer y Lambda* del cuadro de diálogo *Tablas de contingencia: Estadísticos* y se seleccionan las nuevas variables, Tipo e Internet.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	40,523[a]	2	,000
Razón de verosimilitud	55,544	2	,000
Asociación lineal por lineal	39,676	1	,000
N de casos válidos	114		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5,79.

- El estadístico Chi-cuadrado toma el valor 40,523 que permite rechazar la hipótesis de independencia para cualquier nivel de significación.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,596	,000
	V de Cramer	,596	,000
	Coeficiente de contingencia	,512	,000
N de casos válidos		114	

a. No asumiendo la hipótesis nula.

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

- Los valores de los coeficientes Phi, V de Cramer y coeficiente de contingencia de Pearson son superiores a 0,5 y significativos para cualquier nivel. Teniendo en cuenta

que el coeficiente de contingencia para una tabla de dimensiones 3x2 tiene una cota superior inferior a 1, el valor 0,512 indica un grado de asociación moderadamente alto.

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,245	,083	2,534	,011
		TIPO dependiente	,000	,000	,[c]	,[c]
		INTERNET dependiente	,418	,129	2,534	,011
	Tau de Goodman y Kruskal	TIPO dependiente	,157	,038		,000[d]
		INTERNET dependiente	,355	,025		,000[d]

a. No asumiendo la hipótesis nula.

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

c. No se puede efectuar el cálculo porque el error típico asintótico es igual a cero.

d. Basado en la aproximación chi-cuadrado.

- Del análisis direccional se deduce que:

- Tomando la variable Tipo como dependiente, λ Tipo = 0,000, lo cual indica que el conocimiento de los valores de la variable Internet no permite predecir el comportamiento de la variable Tipo.

- Tomando la variable Internet como dependiente, λ Internet=0,418 con un error típico asintótico igual a 0,129. Así pues, el conocimiento de los valores de la variable Tipo permite reducir la incertidumbre en la predicción del comportamiento de la variable Internet en un 41,8%. Esta estimación de λ Internet es significativa para niveles de significación superiores a 0,011.

- El estadístico Tau de Goodman con Internet dependiente toma el valor τ Internet = 0,355 con un error típico de 0,025; este resultado confirma las conclusiones del apartado anterior.

Ejemplo 5:

*Del archivo **Encinf.sav** analice si existe concordancia entre las valoraciones dadas por los alumnos al equipamiento informático del centro, con referencia a las variables*

Dotacion y Software, e indique si es posible establecer una relación de dependencia entre ambas.

Para analizar la concordancia entre variables ordinales se activan las opciones *Correlaciones, Gamma, d de Somers, Tau-b de Kendall y Tau-c de Kendall* del cuadro de diálogo *Tablas de contingencia:Estadísticos*.

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,333	,074	4,428	,000
		dotación dependiente	,333	,074	4,428	,000
		SOFTWARE dependiente	,333	,074	4,428	,000

a. No asumiendo la hipótesis nula.

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

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,333	,074	4,428	,000
	Tau-c de Kendall	,311	,070	4,428	,000
	Gamma	,391	,084	4,428	,000
	Correlación de Spearman	,424	,088	4,769	,000[c]
Intervalo por	R de Pearson	,524	,086	6,281	,000[c]
N de casos válidos		106			

a. No asumiendo la hipótesis nula.

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

c. Basada en la aproximación normal.

- Las medidas simétricas de concordancia entre Dotación y Software indican que existe una asociación moderada y positiva. El coeficiente de correlación de Spearman, que es la medida de concordancia más utilizada con datos ordinales, es 0,424 con un error típico asintótico de 0,088; por tanto, difiere significativamente de 0 para cualquier nivel de significación.

- Las medidas direccionales indican que la concordancia es simétrica ya que coinciden todos los valores del coeficiente d de Somers. Esto quiere decir, que si bien las variables presentan un moderado grado de concordancia no existe entre ellas una relación de dependencia.

Ejemplo 6:

Analice la concordancia entre las variables Mantenimiento y Aulas del archivo ***Encinf.sav****, e indique si puede admitirse una relación de dependencia entre ambas.*

En el cuadro de diálogo *Tablas de contingencia* se seleccionan las variables Mantenimiento y Aulas en las *Filas* y *Columnas*. Con el botón *Estadísticos* se accede al menú donde se activan las medidas de concordancia entre variables ordinales. Los resultados de este análisis son:

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,018	,083	,212	,832
		Mantenimiento dependiente	,017	,081	,212	,832
		Acceso aulas dependiente	,018	,084	,212	,832

a. No asumiendo la hipótesis nula.

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

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,018	,083	,212	,832
	Tau-c de Kendall	,017	,079	,212	,832
	Gamma	,020	,096	,212	,832
	Correlación de Spearman	,028	,106	,282	,778[c]
Intervalo por	R de Pearson	,111	,115	1,127	,262[c]
N de casos válidos		104			

a. No asumiendo la hipótesis nula.

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

c. Basada en la aproximación normal.

Tanto el coeficiente de correlación de Spearman como las restantes medidas simétricas toman valores próximos a 0 y en todos los casos se acepta la hipótesis de que no existe concordancia. A la vista de este resultado no puede suponerse que exista una relación de dependencia entre estas variables; los valores de las medidas direccionales lo confirman.

MEDIDAS DE ASOCIACIÓN PARA VARIABLES CUANTITATIVAS

Para variables cuantitativas, es decir, aquellas que se miden en una escala de intervalo o de razón, las medidas de asociación más utilizadas son la covarianza y el coeficiente de correlación de Pearson. Ambas medidas hacen referencia a un tipo particular de asociación: la asociación lineal.

El análisis conjunto de dos variables cuantitativas puede ir acompañado del análisis unidimensional de cada una de ellas por separado, así como de gráficos que pongan de manifiesto el patrón de comportamiento conjunto de ambas variables.

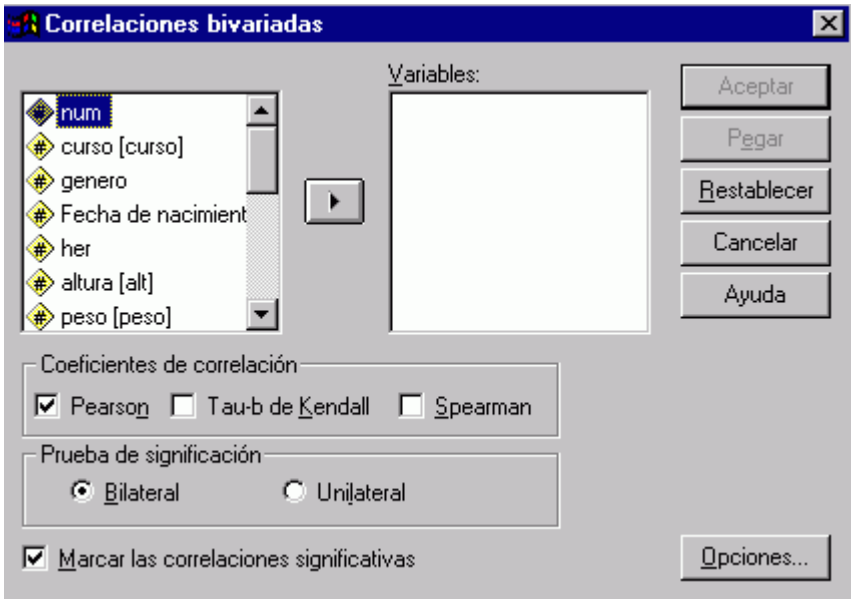
Para realizar el análisis conjunto de dos variables X_1 y X_2 la secuencia es:

Analizar

Correlaciones

Bivariadas

Aparece el siguiente cuadro de diálogo:



Por defecto, están activadas las opciones *Coeficiente de correlación de Pearson*, *Prueba de significación Bilateral y Marcar las correlaciones significativas*. Otras medidas de asociación son los coeficientes *Tau-b de Kendall y Spearman* para variables ordinales, a los que ya se ha hecho referencia en el epígrafe anterior.

COEFICIENTE DE CORRELACIÓN DE PEARSON

El coeficiente de correlacion de Pearson es la principal medida de asociación lineal entre dos variables cuantitativas y se define como:

$$r = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2}\sqrt{\sum_{i=1}^{n}(X_{2i} - \overline{X}_2)^2}}$$

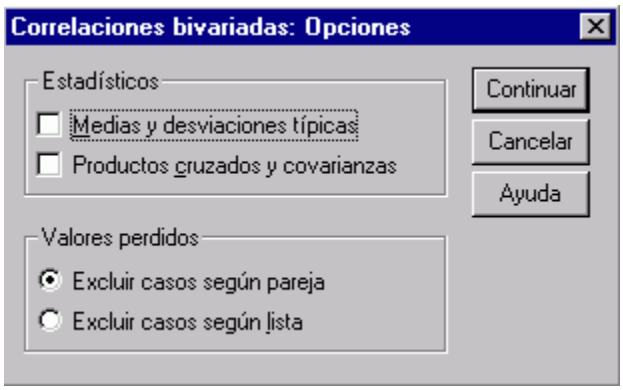
Este coeficiente, cuyo valor no depende de las unidades de medida de las variables, está acotado entre -1 y +1; su signo indica la dirección, positiva o negativa, de la asociación lineal y su valor absoluto la intensidad de la misma. En caso de asociación lineal perfecta toma el valor |1| ; si no hay asociación lineal toma el valor 0, lo cual no implica que no pueda haber asociación de otro tipo.

La prueba de significación del coeficiente de correlación de Pearson puede plantearse:

- *Bilateral* (a doble cola) si se contrasta la hipótesis nula de ausencia de asociación lineal sin especificar de antemano en la hipótesis alternativa la dirección o sentido de la asociación, en caso de que ésta exista.
- *Unilateral* (a una sola cola) si se contrasta la hipótesis nula especificando de antemano en la hipótesis alternativa la dirección de la asociación. Si se desea un contraste unilateral es necesario activar la opción correspondiente.

OPCIONES

Al activar el botón *Opciones* se abre el cuadro de diálogo siguiente.



Este cuadro permite incluir en los resultados los siguientes *Estadísticos*:

- *Medias y desviaciónes típicas* de cada una de las variables análizadas.

- Productos cruzados y covarianzas. La suma de los productos cruzados es el numerador del coeficiente de correlación lineal que dividido por n-1 recibe el nombre de covarianza cuya expresión es:

$$S_{X_1X_2} = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X_1})(X_{2i} - \overline{X_2})}{n-1}$$

La covarianza es una medida de asociación lineal cuyo signo indica la dirección o sentido de la asociación, pero cuyo valor numérico es de difícil interpretación porque depende de las unidades de medida de las variables.

El cuadro de diálogo *Opciones* permite modificar la forma en que se gestionan los valores missing. Por defecto, está activada la opción *Excluir casos según pareja* con la que se calculan los coeficientes de correlación utilizando todos los casos para los que existe información sobre las dos variables. Como alternativa puede activarse la opción *Excluir casos según lista* con la que se calculan los coeficientes de correlación utilizando únicamente los casos para los que se dispone de información sobre todas las variables. Si únicamente se han seleccionado dos variables en el cuadro de diálogo *Correlaciones bivariadas* ambas opciones proporcionan los mismos resultados.

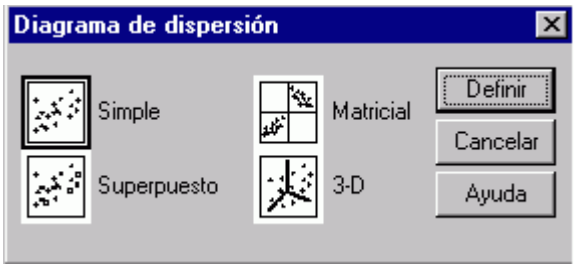
DIAGRAMA DE DISPERSIÓN

La representación gráfica más útil para describir el comportamiento conjunto de dos variables es el diagrama de dispersión o nube de puntos, donde cada caso aparece representado como un punto en el plano definido por las variables X_1 y X_2. Para obtener un diagrama de dispersión la secuencia es:

Gráficos

Dispersión

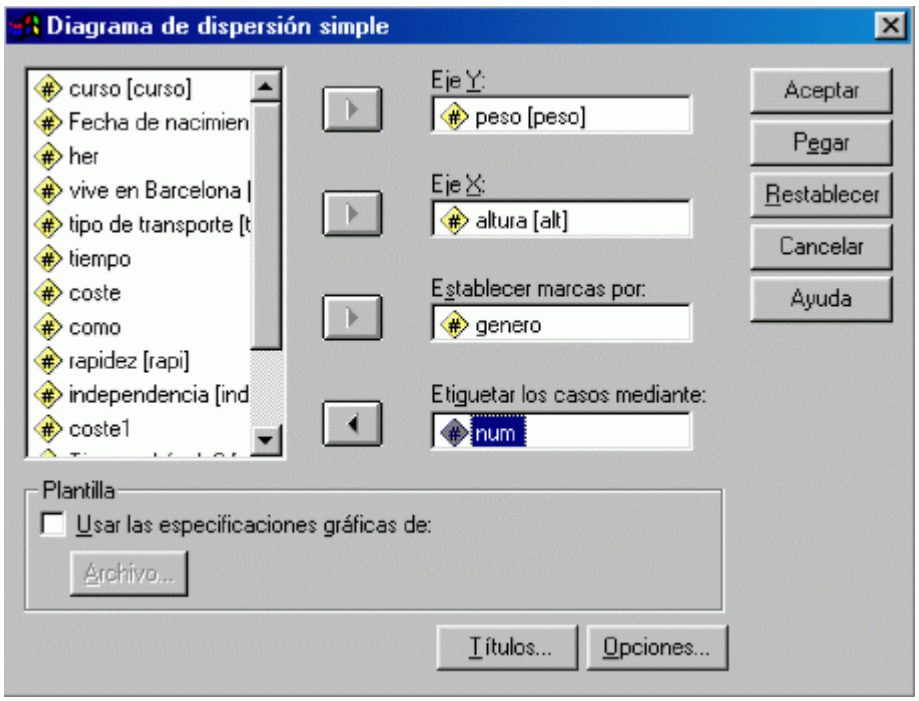
El cuadro de diálogo siguiente:



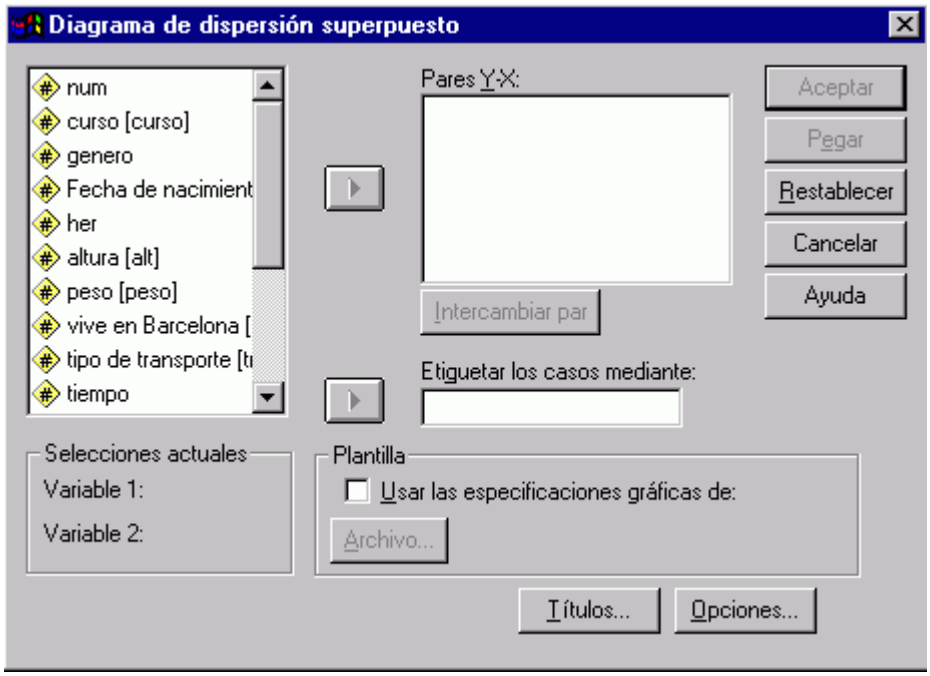
recoge diferentes tipos de diagramas de dispersión. Éstos pueden ser:

1. Simple: si el diagrama sólo recoge el comportamiento simultáneo de dos variables, una definida en el eje X (abscisas) y la otra en el eje Y (ordenadas). Con el botón *Definir* se abre el siguiente cuadro:

- En *Eje X* se selecciona la variable que se considera independiente y en *Eje Y* la dependiente.
- En *Establecer marcas por* puede indicarse alguna variable de control cuyas categorías o valores se representan con un símbolo o color distintivo. Esto permite identificar los puntos pertenecientes a cada categoría y poner de manifiesto si existen comportamientos diferenciados.
- En *Etiquetar los casos mediante* se puede indicar alguna variable cuyos valores se tomarán como etiquetas de los casos. Para visualizar las etiquetas es preciso activar la opción *Mostrar el gráfico* con las etiquetas de caso del cuadro de diálogo *Opciones*.
- El botón *Títulos* ofrece la posibilidad de definir dos líneas de título y un subtítulo, y dos líneas de nota al pie del gráfico.



2. *Superpuestos*: presenta dos o más parejas de variables en un mismo gráfico.

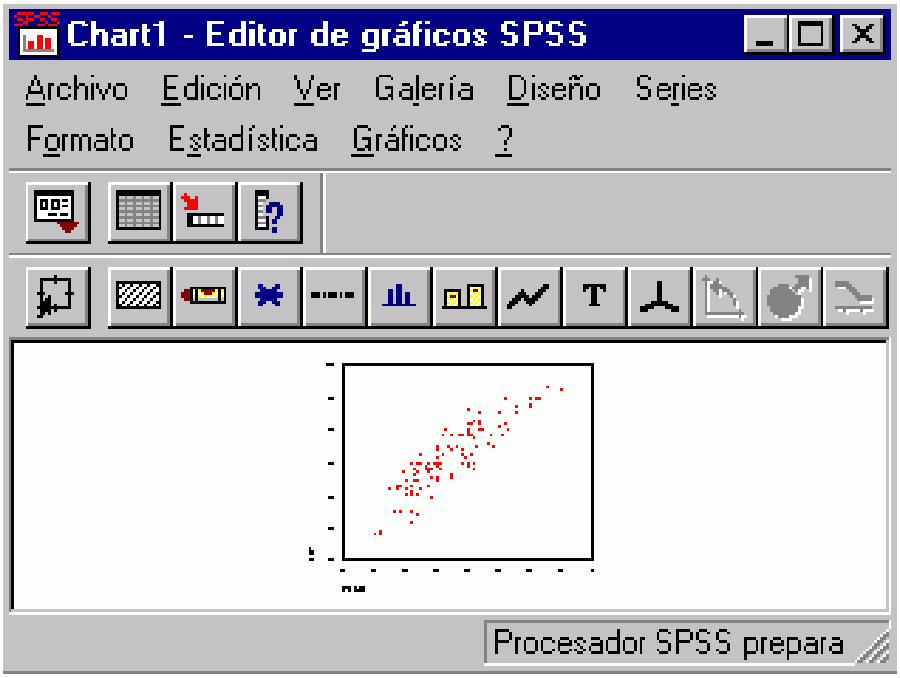


- En *Pares Y-X* se indican las parejas de variables a representar seleccionándolas de dos en dos en la lista de variables. Si se quiere intercambiar X por Y se utiliza el botón *Intercambiar par*.
- *Etiquetar los casos mediante* tiene la misma función que en el diagrama simple.
- Los botones *Titulos y Opciones* ofrecen las mismas posibilidades ya vistas para el diagrama de dispersión simple.

3. *Matricial:* ofrece una matriz de diagramas de dispersión simples de todos los pares y todas las ordenaciones posibles que se pueden formar con las variables seleccionadas. En el cuadro de diálogo que aparece con el botón *Definir* se deben seleccionar las variables cuyos diagramas de dispersión simples aparecerán en la matriz.

4. *3-D*: proporciona en tres dimensiones el diagrama de dispersión de tres variables.

Si el diagrama de dispersión es *Simple o Superpuesto* se puede visualizar con la recta que mejor se ajusta a la nube de puntos. Para ello se edita el gráfico en el visor de resultados haciendo doble clic sobre el mismo.



En la barra de menú del editor de gráficos se activa *Diseño* > *Opciones* y se abre el cuadro de diálogo:

Se selecciona *Ajustar línea* > *Total*. En *Opciones de ajuste* se puede elegir el método de ajuste deseado entre: *Regresión lineal* (activado por defecto), *Regresión cuadrática, Regresión cúbica y Minsce*. También es posible incluir en el diagrama de dispersión una línea paralela al eje de abscisas que pasa por la media de la variable Y con la opción *Línea de referencia* para la media en *Y* > *Total.*

Cuando el diagrama recoge un gran número de observaciones algunos puntos representan a más de un caso ya que estos se superponen . Con la opción *Girasoles* > *Mostrar girasoles* cada punto aparece con tantas rayas o 'pétalos' como casos

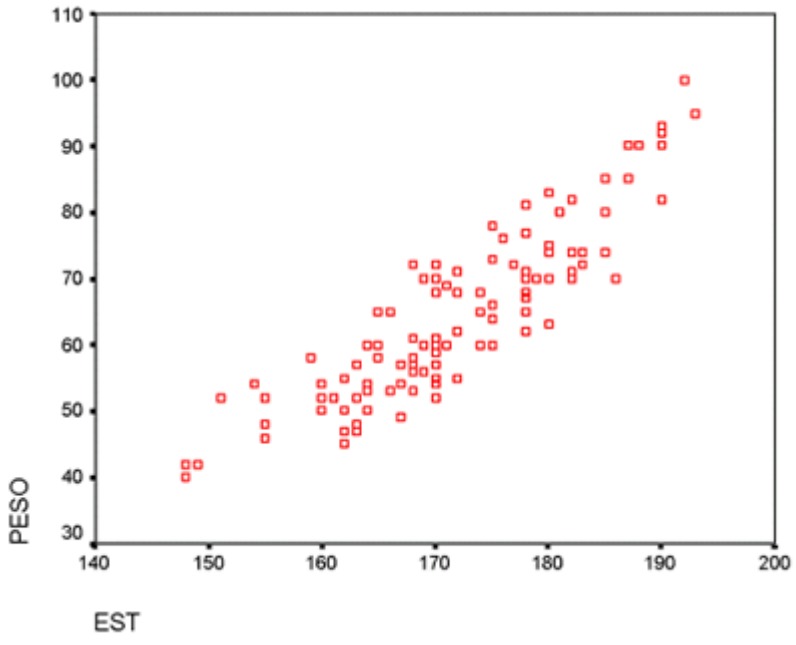
representa. Ésta es una forma gráfica de indicar cuantos casos están representados por un punto.

EJEMPLOS

Con referencia a las variables Peso y Est (estatura) del archivo ***Encinf.sav*** *comprobar gráfica y analíticamente la existencia de una relación lineal entre ellas.*

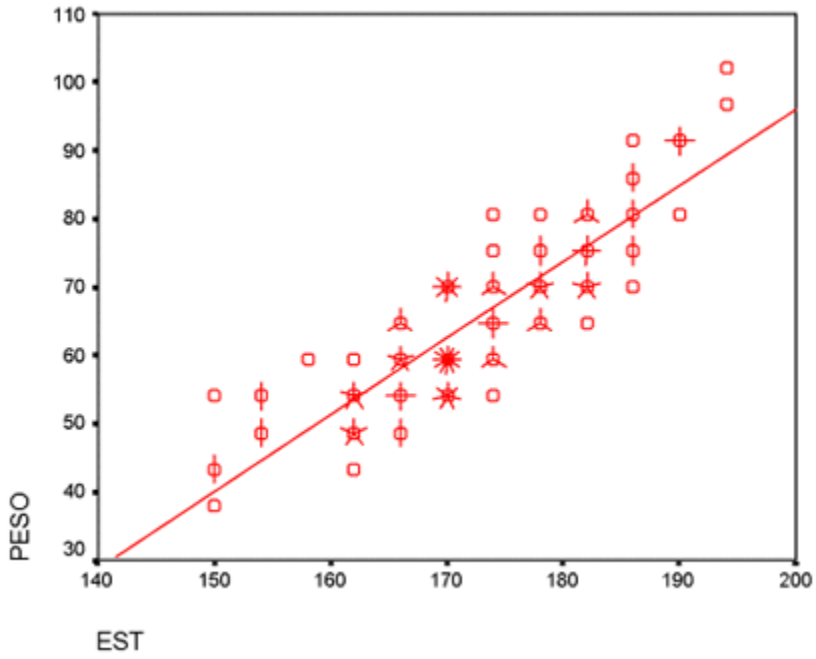
1. La representación gráfica que permite comprobar la existencia de relación lineal entre dos variables es el diagrama de dispersión y la medida analítica adecuada es el coeficiente de correlación lineal.

Con la secuencia *Gráficos* > *Dispersión* > *Simple* > *Definir* se abre el cuadro de diálogo *Diagrama de dispersion simple*. Se seleccionan en el *Eje Y* la variable Peso y en el *Eje X* la variable Est, y se obtiene el siguiente gráfico:



Como se observa en el gráfico ambas variables presentan una relación lineal positiva; es decir, a medida que aumenta el valor de la variable Est aumenta también el valor de la variable Peso.

Si se desea ver la nube de puntos con la línea de mejor ajuste superpuesta, y que los casos iguales o muy próximos entre si queden representados por un sólo punto, se edita el gráfico haciendo doble clic sobre el mismo. En el editor de gráficos se selecciona *Diseño* > *Opciones y en el cuadro Opciones del diagrama de dispersión* se activa *Mostrar girasoles y Ajustar línea* > *Total*.



Para cuantificar el grado de asociación lineal entre las variables la medida adecuada es el coeficiente de correlación lineal de Pearson. Con la secuencia *Analizar* > *Correlaciones* > *Bivariadas* se abre un cuadro de diálogo donde se seleccionan las variables Peso y Est, y con las opciones *Coeficiente de correlación* > *Pearson* se obtiene la siguiente matriz de correlaciones.

Correlaciones

		EST	PESO
EST	Correlación de Pearson		
	Sig. (bilateral)		
	N		
PESO	Correlación de Pearson	,883**	
	Sig. (bilateral)	,000	
	N	114	

**. La correlación es significativa al nivel 0,01

El valor de r=0,883 es positivo y elevado, así como significativo para cualquier nivel, con lo cual se confirma la impresión proporcionada por el gráfico acerca de la existencia de asociación lineal entre las variables

Cuarta unidad

.....

La Inferencia Estadística

4.1 LOS MÉTODOS PARAMÉTRICOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo realizará adecuadamente una inferencia estadística de la media muestral y a la media poblacional, a través de los métodos paramétricos.

LOS MÉTODOS PARAMÉTRICOS

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.

Los métodos paramétricos de la inferencia estadística se pueden dividir, básicamente, en dos: métodos de estimación de parámetros y métodos de contraste de hipótesis. Ambos métodos se basan en el conocimiento teórico de la distribución de probabilidad del estadístico muestral que se utiliza como estimador de un parámetro.

La estimación de parámetros consiste en asignar un valor concreto al parámetro o parámetros que caracterizan la distribución de probabilidad de la población. Cuando se estima un parámetro poblacional, aunque el estimador que se utiliza posea todas las propiedades deseables, se comete un error de estimación que es la diferencia entre la estimación y el verdadero valor del parámetro. El error de estimación es desconocido por lo cual es imposible saber en cada caso cual ha sido la magnitud o el signo del error; para valorar el grado de precisión asociado con una estimación puntual se parte de dicha estimación para construir un intervalo de confianza. En síntesis, un intervalo de confianza está formado por un conjunto de valores numéricos tal que la probabilidad de que éste contenga al verdadero valor del parámetro puede fijarse tan grande como se quiera. Esta probabilidad se denomina grado de confianza del intervalo, y la amplitud deéste constituye una medida del grado de precisión con el que se estima el parámetro.

Los métodos de contraste de hipótesis tienen como objetivo comprobar si determinado supuesto referido a un parámetro poblacional, o a parámetros análogos de dos o más poblaciones, es compatible con la evidencia empírica contenida en la muestra. Los supuestos que se establecen respecto a los parámetros se llaman hipótesis paramétricas. Para cualquier hipótesis paramétrica, el contraste se basa en establecer

un criterio de decisión, que depende en cada caso de la naturaleza de la población, de la distribución de probabilidad del estimador de dicho parámetro y del control que se desea fijar a priori sobre la probabilidad de rechazar la hipótesis contrastada en el caso de ser ésta cierta.

En todo contraste intervienen dos hipótesis. La hipótesis nula (Ho) es aquella que recoge el supuesto de que el parámetro toma un valor determinado y es la que soporta la carga de la prueba. La decisión de rechazar la hipótesis nula, que en principio se considera cierta, está en función de que sea o no compatible con la evidencia empírica contenida en la muestra. El contraste clásico permite controlar a priori la probabilidad de cometer el error de rechazar la hipótesis nula siendo ésta cierta; dicha probabilidad se llama nivel de significación del contraste ( ) y suele fijarse en el 1%, 5% o 10%.

La proposición contraria a la hipótesis nula recibe el nombre de hipótesis alternativa (H1) y suele presentar un cierto grado de indefinición: si la hipótesis alternativa se formula simplemente como 'la hipótesis nula no es cierta', el contraste es bilateral o a dos colas; por el contrario cuando se indica el sentido de la diferencia, el contraste es unilateral o a una sola cola.

Cuando se realiza un contraste con el SPSS no se fija el nivel de significación deseado, el programa calcula el valor-p o significación asintótica, que es la probabilidad de que el estadístico de prueba tome un valor igual o superior al muestral bajo el supuesto de que la hipótesis nula es cierta. Por tanto, si el valor-p es menor o igual que el nivel de significación deseado se rechazará Ho.Un valor-p próximo a cero indica que se rechazará la Ho para cualquier nivel de significación.

MEDIA POBLACIONAL

Para poblaciones normales o aproximadamente normales el intervalo de confianza para la media poblacional está centrado en la media muestral; siendo sus límites, superior e inferior, donde es el valor crítico correspondiente al grado de confianza de la distribución t de Student con n-1 grados de libertad. En la práctica si n es moderadamente grande el valor crítico es igual a 1,64 para un intervalo del 90% de confianza, 1,96 para el 95%, o 2,58 para el 99%.

Para obtener el intervalo de confianza para la media la secuencia es:

Analizar

Estadísticos Descriptivos

Explorar

DE MEDIAS POBLACIONALES $\mu_1 - \mu_2$

En ocasiones interesa definir un intervalo de valores tal que permita establecer cuales son los valores mínimo y máximo aceptables para la diferencia entre las medias de dos poblaciones. Pueden darse dos situaciones según las muestras sean o no independientes; siendo en ambos casos condición necesaria que las poblaciones de origen sean normales o aproximadamente normales:

- MUESTRAS INDEPENDIENTES

Si puede suponerse que las varianzas de ambas poblaciones son iguales, el intervalo de confianza para la diferencia de medias poblacionales está centrado en la diferencia de las medias muestrales, siendo sus límites superio e inferior:

$t^{\alpha}/2$ es el valor crítico correspondiente al grado de confianza *1-*α de la distribución t de Student con n1+ n2-2 grados de libertad y $S = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$ es una estimación de la desviación típica común a ambas poblaciones obtenida a partir de las varianzas de las dos muestras. En la práctica si n1 y n2 son moderadamente grandes, el valor crítico

$t^{\alpha}/2$ se aproxima, como ya se ha visto anteriormente, a los valores de la distribución normal.

Si las varianzas poblacionales no pueden suponerse iguales los límites del intervalo de confianza son:

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

El valor crítico $t^{\alpha}/2$ corresponde a una distribución t cuyos grados de libertad se calculan en base a ambos tamaños muestrales y a las desviaciones típicas de cada grupo según la corrección propuesta por Dixon y Massey:

$$g.l.=\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2\left(\frac{1}{n+1}\right)+\left(\frac{S_2^2}{n_2}\right)^2\left(\frac{1}{n+1}\right)}-2$$

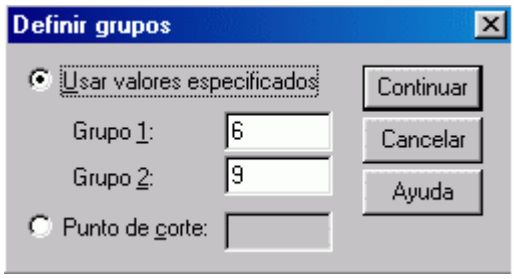
Para obtener el intervalo de confianza en ambos casos la secuencia es:

Analizar

Comparar medias

Prueba T para muestras independientes

En el cuadro de diálogo hay que seleccionar en *Contrastar variables* la variable objeto de análisis e indicar la *Variable de agrupación* junto con el criterio para *Definir grupos* (las dos poblaciones).



Los grupos pueden definirse en función de una variable cuantitativa o de una cualitativa. Si la variable de agrupación presenta sólo dos valores o modalidades, entonces se debe seleccionar *Usar valores especificados* e indicar la modalidad que define el grupo 1 y la del grupo 2. Si la variable tiene más de 2 valores o modalidades se elige la opción *Punto de corte* indicando el valor de la variable que induce una

partición en dos grupos, uno de los cuales estará formado por todos los casos con valores menores que el especificado y el otro por el resto de casos.

Al aceptar se obtienen:

- resultados de la prueba de Levene para contrastar la igualdad de varianzas $(H_0{:}\sigma_1^2 = \sigma_2^2)$ *

- resultados de la prueba T para contrastar la igualdad de medias $(H_0{:}\mu_1 = \mu_2)$

- intervalo de confianza para la diferencia de medias al 95% por defecto.

Si se quiere cambiar el grado de confianza del intervalo, antes de aceptar hay que modificarlo con el botón *Opciones*.

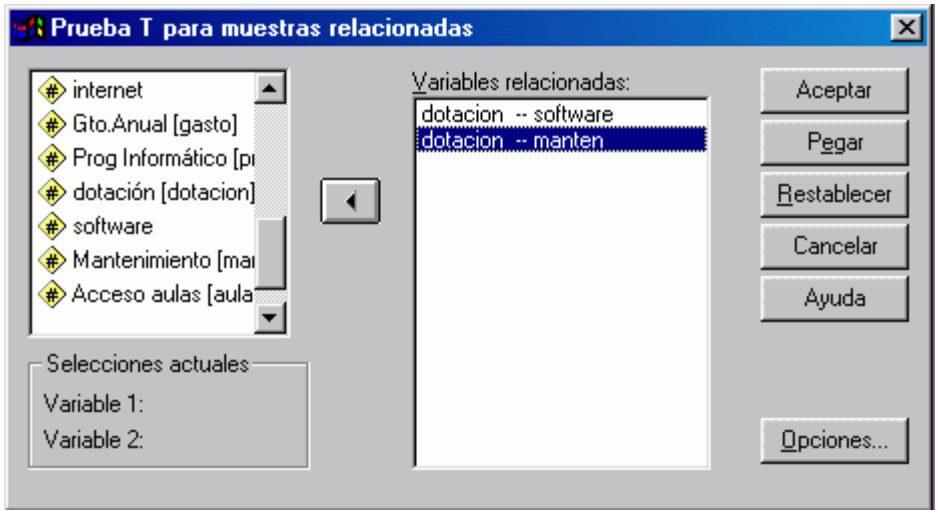
- MUESTRAS DEPENDIENTES. En este caso las muestras están formadas por parejas de valores, uno de cada población y el estadístico se obtiene a partir de las diferencias de los valores de las dos variables correspondientes a cada caso o d*i* que se define como d*i*= xi-*yi*.

Para contrastar la hipótesis de igualdad de medias y obtener el intervalo de confianza la secuencia es:

Analizar

Comparar medias

Prueba T para muestras independientes



Las variables se deben seleccionar por parejas. Haciendo clic sobre las variables de la lista aparecen sus nombres en el cuadro *Selecciones actuales*; una vez seleccionadas las dos variables se trasladan al recuadro *Variables relacionadas* de la forma habitual. En cada sesión se pueden seleccionar tantos pares de variables como medias se quieran comparar.

Al aceptar se obtienen para cada par de variables, los siguientes resultados:

- Media, desviación tipo y error típico de la media de las d*i.*

- Intervalo de confianza, por defecto al 95%, para la diferencia de medias poblacionales que viene dado por:

$(\bar{d} \pm t_{\alpha/2}\frac{S_d}{\sqrt{n}})$ donde $t_{\alpha/2}$ es el valor de la distribución t de Student con n-1 grados de libertad que deja por encima una probabilidad de α/2.

- Estadístico t del contraste $H_0{:}\mu_1 = \mu_2$ y $H_1{:}\mu_1 \neq \mu_2$ (contraste de dos colas).

EJEMPLO

Ejemplo 1

Con los datos de una encuesta obtener la estimación puntual y los intervalos de confianza del 95 y del 99% para la media de la población de la variable Coste.

En el cuadro de diálogo *Explorar*, que se obtiene con la secuencia *Analizar* > *Estadísticos descriptivos* > *Explorar*, se selecciona como variable dependiente la variable Coste. En Estadísticos comprobamos que está activada la opción *Descriptivos* y que el intervalo para la media definido es el del 95%.

Al aceptar se obtiene el siguiente cuadro de resultados:

Descriptivos

			Estadístico	Error típ.
COSTE	Media		5236,40	365,97
	Intervalo de confianza para la media al 95%	Límite inferior	4511,34	
		Límite superior	5961,46	
	Media recortada al 5%		4813,81	
	Mediana		4000,00	
	Varianza		15268788	
	Desv. típ.		3907,53	
	Mínimo		0	
	Máximo		20000	
	Rango		20000	
	Amplitud intercuartil		2387,50	
	Asimetría		2,076	,226
	Curtosis		4,836	,449

La estimación puntual del valor esperado del coste es 5236,40 Pta. Esta estimación tiene un error típico de 365,97. Los límites inferior y superior del intervalo de confianza del 95% son 4511,34 y 5951,46, respectivamente. Este resultado se interpreta como que de los intervalos obtenidos con este método el 95% contendrán el verdadero valor esperado del coste. Una medida del grado de precisión con el que se está estimando el valor esperado es la amplitud del intervalo, que en este caso es igual a 1450,12 y la mitad de la amplitud, que es 725,06, es el error máximo de estimación que puede garantizarse con una probabilidad de 0,95. Este error máximo es igual a $t_{\alpha/2}\frac{S}{\sqrt{n}}$, donde $t^{\alpha}/2$, es el valor crítico para α=0,05 de la distribución t e Student, en este caso con 113 grados de libertad, y $\frac{S}{\sqrt{n}}$ es el error típico de la estimación.

Para obtener el intervalo del 99% de confianza modificamos el valor del grado de confianza en el cuadro Explorar: Estadísticos fijándolo en el 99%.

Descriptivos

			Estadístico	Error tip.
COSTE	Media		5236,40	365,97
	Intervalo de confianza para la media al 99%	Límite inferior	4277,54	
		Límite superior	6195,27	
	Media recortada al 5%		4813,81	
	Mediana		4000,00	
	Varianza		15268788	
	Desv. tip.		3907,53	
	Mínimo		0	
	Máximo		20000	
	Rango		20000	
	Amplitud intercuartil		2387,50	
	Asimetría		2,076	,226
	Curtosis		4,836	,449

Los límites del intervalo de confianza del 99% son 4277,54 y 6195,27; la confianza de que este intervalo contenga el verdadero valor esperado del coste es 0,99. La amplitud de este intervalo es 2217,73 que es mayor que la amplitud del intervalo del 95%, por lo tanto, 1108,865, es el error máximo de estimación que puede garantizarse con una probabilidad de 0,99. Como puede verse, a medida que aumenta el grado de confianza del intervalo disminuye la precisión de la estimación.

Ejemplo 2.

Para la misma variable Coste verificar si se puede aceptar el supuesto de que el valor esperado del Coste es superior a 6000.

Con la secuencia *Analizar* > *Comparar medias* > *Prueba T para una media* se abre el cuadro de diálogo *Prueba T para una muestra* en el cual se selecciona la variable Coste y se indica como *Valor de prueba* 6000. Esto quiere decir que las hipótesis que se están contrastando son $H_0{:}\mu = 6000$ frente a $H_1{:}\mu > 6000.$ Se trata por tanto de un contraste a una sola cola.

Prueba para una muestra

	Valor de prueba = 6000					
					Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
COSTE	-2,086	113	,039	-763,60	-1488,66	-38,54

El estadístico de prueba $t=\frac{\overline{X}-6000}{S/\sqrt{n}}$ toma el valor t=-2,086, que en las tablas de la distribución t de Student con 113 grados de libertad deja por debajo un área de 0,0195. Esto quiere decir que se rechaza la hipótesis nula en favor de la alternativa para niveles de significación superiores a 0,0195. Obsérvese que 0,0195 es la mitad del nivel de significación para la prueba de dos colas que aparece en el cuadro de resultados.

Por otra parte si las hipótesis hubieran sido $H_0{:}\mu = 6000,\ H_1{:}\mu \neq 6000$ se rechazaría la hipótesis nula en favor de la alternativa para niveles de significación superiores a 0,039. El intervalo del 95% de confianza para la media calculado en el apartado anterior no contenía el valor 6000; lo que equivale a decir que para un nivel de significación del 5% se rechaza la hipótesis nula. Por el contrario, el intervalo del 99% contenía el valor 6000 y, por lo tanto, para un nivel de significación del 1% no se rechazaría la hipótesis nula.

Ejemplo 3.

Verificar si existe diferencia significativa entre el coste esperado en transporte de los alumnos que viven en Barcelona y el de los que viven fuera.

Con la secuencia *Analizar> Comparar medias > Prueba T para muestras independientes* se abre el cuadro de diálogo *Prueba T para muestras independientes* en el cual se selecciona la variable Coste y se indica como *Variable de agrupación* Resid. En la opción *Definir grupos* se asigna al *Grupo 1* el valor 1 (vive en Lima) y al *Grupo 2* el valor 2 (no vive en Lima). Aceptando se obtienen entre otros los siguientes resultados:

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
COSTE	Se han asumido varianzas iguales	37,671	,000	-4,459	112	,000
	No se han asumido varianzas iguales			-3,750	50,997	,000

Las hipótesis que se están contrastando son frente $H_0{:}\mu_1 = \mu_2$ frente a $H_1{:}\mu_1 \neq \mu_2$. Para realizar este contraste previamente se debe comprobar si es aceptable la hipótesis de varianzas poblacionales iguales para los dos grupos $H_0{:}\sigma_1^2 = \sigma_2^2$. El estadístico F de la prueba de Levene* no permite aceptar la igualdad de varianzas poblacionales, por lo cual el valor del estadístico de prueba es t=-3,750 que para cualquier nivel de significación lleva a rechazar la hipótesis de igualdad de medias. El signo negativo del estadístico t indica que el coste del transporte es significativamente superior para los que viven fuera de Barcelona.

Ejemplo 4.

Con los datos de una encuesta contrastar si existe diferencia significativa entre las puntuaciones medias asignadas en cuanto a una variable cualitativa .

Las puntuaciones que se quiere comparar han sido generadas dos a dos por los mismos individuos; se trata por tanto del caso de muestras relacionadas. Las hipótesis que se contrastan son $H_0{:}\mu_1 = \mu_2$ frente a $H_1{:}\mu_1 \neq \mu_2$.

Con la secuencia *Analizar > Comparar medias >Prueba T para muestras relacionadas* se abre el cuadro de diálogo en el cual se selecciona la pareja de variables. Al aceptar se obtienen los siguientes resultados:

Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	Intervalo de confianza para la diferencia Inferior	Superior			
dotación - SOFTWARE	-1,12	1,95	,19	-1,50	-,75	-5,93	105	,000

El análisis sólo ha considerado los casos que no presentan ningún valor missing en el par de puntuaciones, quedando únicamente 106 casos válidos de los 114.

El promedio de las diferencias entre las puntuaciones asignadas a la dotación y al software es de -1,12 con un error típico igual a 0,19. El estadístico de prueba t es igual a -5,93 y se distribuye según una t de Student con 105 grados de libertad. Con este valor de t se rechaza la hipótesis nula para cualquier nivel de significación. Los resultados proporcionan también el intervalo de confianza para la diferencia de las dos medias poblacionales con el 95% de nivel de confianza. Como puede observarse el

intervalo no contiene el valor 0, de lo que se deduce también que no se puede aceptar que las puntuaciones medias sean significativamente iguales.

PREGUNTAS

- **¿Cómo se infiere a la Media Poblacional?**
- **¿Cómo se elabora un análisis de datos con la T de student Pareada y para muestras independientes?**

4.2 LOS MÉTODOS NO PARAMÉTRICOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo realizará adecuadamente una inferencia estadística de la proporción poblacional, a través de los métodos no paramétricos.

PROPORCIÓN POBLACIONAL π

En poblaciones dicotómicas con una proporción π de éxitos el estimador puntual del parámetro π es la proporción muestral de éxitos, p, que coincide con la media de la muestra cuando se codifica como 1 la característica que se considera como éxito y 0 la que se considera no éxito. A partir de un tamaño muestral moderadamente grande el estadístico p tiene una distribución aproximadamente normal. El intervalo de confianza para la proporción poblacional está centrado en la proporción muestral; siendo sus límites superior e inferior $p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ donde $z^{\alpha}/2$ es el valor crítico correspondiente al grado de confianza 1- α de la distribución normal tipificada y $\sqrt{\frac{p(1-p)}{n}}$ es el error típico de la proporción.

Para obtener el intervalo de confianza y contrastar hipótesis sobre la proporción una alternativa consiste en tratar a la proporción π como la media poblacional de una variable dicotómica codificada como se ha descrito anteriormente (éxito=1, no éxito=0) y la secuencia es:

- Para el intervalo de confianza:

Analizar

Estadísticos Descriptivos

Explorar

- Para contrastar la hipótesis nula $H_0{:}\pi = \pi_0{:}$

Analizar

Comparar medias

Prueba T para una muestra

Utilizando este criterio los resultados numéricos no coinciden exactamente con los que se obtendrían aplicando la expresión del error típico de la proporción; no obstante la

discrepancia es despreciable si el número de observaciones es suficientemente grande.

Otras alternativas para realizar este contraste son de naturaleza no paramétrica.

PRUEBA BINOMIAL

La prueba binomial prueba binomial analiza variables dicotómicas y compara las frecuencias observadas en cada categoría con las que cabría esperar según una distribución binomial de parámetro π especificado en la hipótesis nula. El nivel de significación crítico de esta prueba indica la probabilidad de obtener una discrepancia igual o superior a la observada a partir de la muestra si la distribución es la postulada por la hipótesis nula.

El nivel de significación crítico (bilateral) de este contraste debe interpretarse como:

$P(|X - n\pi_0| \geq X_0 / X \sim B(n, \pi_0))$ el número de éxitos en la muestra.

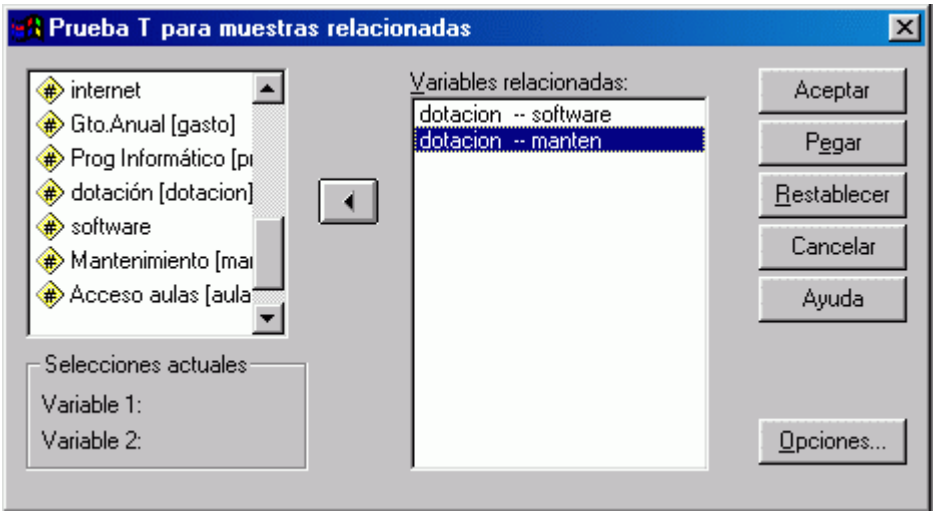
Cuando n es suficientemente grande se calcula esta probabilidad aproximando la distribución binomial a la normal con corrección de continuidad.

La secuencia para realizar este contraste es:

Analizar

Pruebas no paramétrica

Binomial



Se especifica la variable en *Contrastar variables*. Se indica la proporción postulada en la hipótesis nula en *Contrastar proporción*. Si la variable es dicotómica se mantiene activada la opción *Obtener de los datos*. Si la variable no es dicotómica en *Punto de corte* se indica el valor de corte de forma que los inferiores o iguales se agrupan en la primera categoría y el resto en la otra.

El botón *Opciones* permite obtener estadísticos de resumen y modificar el tratamiento de los valores missing.

PRUEBA CHI-CUADRADO

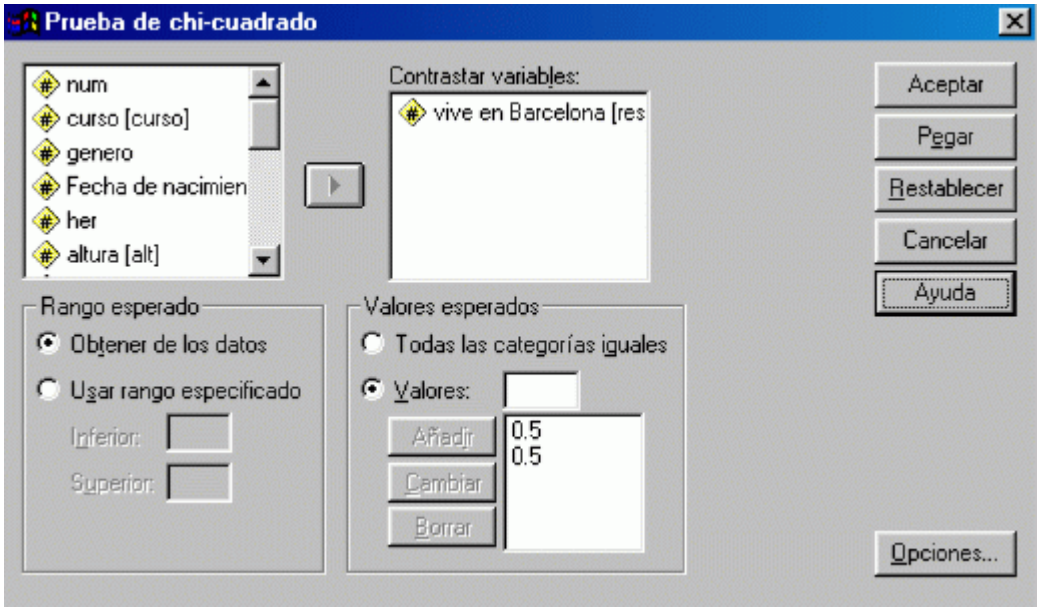
Otra alternativa de naturaleza no paramétrica se basa en el valor del estadístico Chi-cuadrado. Esta prueba compara la frecuencia observada para cada categoría (Oi) con la frecuencia teórica o esperada (Ei) bajo el supuesto de que la proporción de éxitos es la postulada por la hipótesis nula. Para una muestra de tamaño *n* la frecuencia esperada se calcula: $E_0 = \pi_0 n$ para los éxitos y $E_1 = (1-\pi_0)n$ para los no éxitos. El estadístico de prueba Chi-cuadrado se define $\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$ y presenta una distribución Chi-cuadrado con 1 grado de libertad.

La secuencia es:

Analizar

Pruebas no paramétricas

Chi-cuadrado



En el cuadro de diálogo se indica, en la casilla *Contrastar variables* la variable sobre la que se realiza el contraste; obsérvese que en *Rango esperado* está activada la opción *Obtener de los datos*, que toma tantas categorías como valores distintos presenta la variable. Si la hipótesis nula es $H_0{:}\pi = 0,5$ los valores esperados serán *Todas las categorías iguales*, en cualquier otro caso se pueden introducir los valores (frecuencias esperadas absolutas o relativas) con la opción *Valores.*

EJEMPLO

Con los datos de una encuesta estimar la proporción mediante un intervalo de confianza del 90% y contrastar la hipótesis de que son mayoría.

El primer paso es comprobar si la variable Resid está correctamente codificada, es decir, presenta valor 1 cuando el alumno es residente en Barcelona y 0 en otro caso. En la base de datos se ve que la codificación de la variable Resid no es la adecuada, por lo tanto, en primer lugar habrá que recodificar la variable. La secuencia es *Transformar* > *Recodificar* > *En la misma variable*. En el cuadro de diálogo se selecciona la variable Resid y se definen los *Valores antiguos y nuevos* asignando al v*alor antiguo 2 el valor nuevo 0*. En el editor de datos se puede etiquetar el valor 0 como 'no' haciendo doble clic sobre el título de la variable Resid.

- Una aproximación del intervalo de confianza se obtiene con la secuencia *Analizar* > *Estadísticos descriptivos* > *Explorar*. En el cuadro de diálogo se selecciona como variable dependiente la variable Resid y en *Estadísticos* se modifica el grado de confianza fijándolo en el 90%. Los resultados que se obtienen al aceptar son:

Descriptivos

			Estadístico	Error típ.
vive en Barcelona	Media		,61	4,58E-02
	Intervalo de confianza para la media al 90%	Límite inferior	,54	
		Límite superior	,69	
	Media recortada al 5%		,63	
	Mediana		1,00	
	Varianza		,239	
	Desv. típ.		,49	
	Mínimo		0	
	Máximo		1	
	Rango		1	
	Amplitud intercuartil		1,00	
	Asimetría		-,475	,226
	Curtosis		-1,807	,449

Como se observa la media igual a 0,61 indica que el 61% de los elementos de la muestra residen en Barcelona. A partir de esta estimación puntual de π y de la estimación de su error típico se obtiene el intervalo de confianza al 90% cuyo límites inferior y superior son 0,54 y 0,69, respectivamente.

- Para contrastar las hipótesis $H_0{:}\pi = 0,5$ y $H_1{:}\pi > 0,5$ considerando la proporción como media la secuencia es *Analizar* > *Comparar medias* > *Prueba T para una muestra*. Se selecciona la variable Resid y se indica como *Valor de prueba* 0,5, los resultados del contraste se recogen en el siguiente cuadro:

Prueba para una muestra

	Valor de prueba = 0.5					
					Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
vive en Barcelona	2,490	113	,014	,11	2,33E-02	,20

Como el contraste es a una sola cola el nivel de significación crítico es la mitad de 0,014, es decir, 0,007. Para niveles de significación superiores a 0,007 se rechaza la hipótesis nula. Así pues, para los niveles de significación habituales se acepta que la mayoría de alumnos viven en Barcelona (π >0,5).

- Utilizando la prueba binomial para contrastar las hipótesis $H_0{:}\pi = 0,5$ y $H_1{:}\pi > 0,5$ la secuencia es *Analizar* > *Pruebas no paramétricas* > *Binomial.*

Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (bilateral)
vive en Barcelona	Grupo 1	si	70	,61	,50	,019[a]
	Grupo 2	no	44	,39		
	Total		114	1,00		

a. Basado en la aproximación Z.

Si la hipótesis nula es cierta, el número de residentes observado (70) proviene de una población binomial de parámetros n=114 y π =0,5 y por tanto con valor esperado 57. En tal caso:

$$P(|X - 57| \geq 70) = P(X \leq 13) + P(X \geq 70) = 0,019$$

Por lo tanto, $P(X \geq 70)$ = 0,0095 que es el nivel de significación crítico del contraste a cola superior, en consecuencia se rechaza la hipótesis nula.

- Para contrastar las hipótesis $H_0{:}\pi = 0,5$ y $H_1{:}\pi > 0,5$ aplicando la prueba Chi-cuadrado la secuencia es *Analizar* > *Pruebas no paramétricas* > *Chi-cuadrado*.

vive en Barcelona

	N observado	N esperado	Residual
si	70	57,0	13,0
no	44	57,0	-13,0
Total	114		

Estadísticos de contraste

	vive en Barcelona
Chi-cuadrado[a]	5,930
gl	1
Sig. asintót.	,015

a. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 57,0.

Como el nivel de significación asintótico es 0,015 la hipótesis nula no se rechaza para los niveles de significación inferiores al 1,5 %.

DIFERENCIA DE PROPORCIONES $\pi_1 - \pi_2$

El estadístico de prueba que permite contrastar $H_0{:}\pi_1 = \pi_2$ frente a $H_1{:}\pi_1 \neq \pi_2$ a partir de dos muestras aleatorias e independientes es $Z = \frac{p_1 - p_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$ siendo p la estimación de π obtenida del total de observaciones.

Si se consideran las proporciones como medias y se aplica la prueba t utilizada para comparar medias poblacionales los resultados no son fiables ya que la estimación del error típico que realiza el programa no coincide con la del estadístico de prueba. Para resolver el problema con el programa SPSS se deberá cruzar la variable analizada con

la que define los grupos (obtener la tabla de contingencia) y realizar el contraste de independencia Chi-cuadrado.

El estadístico de prueba Chi-cuadrado se define: $\chi^2 = \sum_{j=1}^{2} \sum_{i=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ y presenta una distribución Chi-cuadrado con (I-1)(J-1) grados de libertad. Las Eij se calculan multiplicando las frecuencias marginales y dividendo el producto por n. Estas Eij son estimaciones de las frecuencias absolutas que cabría esperar en cada casilla bajo el supuesto de que la proporción de éxitos es la obtenida a partir del total de observaciones muestrales sin considerar diferencias entre los dos grupos.

La secuencia es:

Analizar

Estadísticos Descriptivos

Tablas de contingencia

En el cuadro de diálogo se indica la variable que se quiere contrastar (filas), la variable que define los dos grupos (columnas) y se selecciona la opción Chi-cuadrado en Estadísticos.

EJEMPLO

*Con referencia a la encuesta **Enctrans.sav** se quiere comprobar si la proporción de alumnos con vehículo difiere significativamente entre los grupos definidos según el género.*

La hipótesis nula del contraste es $H_0{:}\pi_1 = \pi_2$; siendo π_1 la proporción poblacional de hombres con vehículo y π_2 la proporción poblacional de mujeres con vehículo.

Con la secuencia *Analizar* > *Estadísticos Descriptivos* > *Tablas de contingencia* se accede al cuadro de diálogo donde se indica que la variable a contrastar es Vehículo y que la variable de agrupación es el Género, y se selecciona la opción *Chi-cuadrado en Estadísticos*. Al aceptar se obtiene el siguiente cuadro de resultados.

Tabla de contingencia Tiene vehículo? * GENERO

			GENERO		
			hombre	mujer	Total
Tiene vehículo?	No	Recuento	33	42	75
		Frecuencia esperada	35,5	39,5	75,0
		% del total	28,9%	36,8%	65,8%
	Sí	Recuento	21	18	39
		Frecuencia esperada	18,5	20,5	39,0
		% del total	18,4%	15,8%	34,2%
Total		Recuento	54	60	114
		Frecuencia esperada	54,0	60,0	114,0
		% del total	47,4%	52,6%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	,998[b]	1	,318		
Corrección de continuidad[a]	,642	1	,423		
Razón de verosimilitud	,998	1	,318		
Estadístico exacto de Fisher				,331	,212
Asociación lineal por lineal	,989	1	,320		
N de casos válidos	114				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 18,47.

Si es cierto que la proporción de propietarios de vehículo es la misma en los dos grupos, $\pi_1 = \pi_2 = \pi$, la estimación de π es la proporción de propietarios de vehículo para el total de alumnos de la muestra, es decir, 39/114=0,3421. La frecuencia esperada de hombres con vehículo se obtendrá multiplicando esta proporción por el total de hombres en la muestra, o sea, 0,3421·54=18,5; y de la misma forma se obtendrá la frecuencia esperada de mujeres con vehículo: 0,3421·60=20,5 (veáse que estas frecuencias esperadas coinciden con las que cabría esperar en el caso de que las variables Género y Vehículo fueran independientes).

El estadístico Chi-cuadrado toma el valor 0,998 y el nivel de significación crítico es 0,318, por lo tanto no se rechaza la hipótesis nula para los niveles de significación habituales y se puede aceptar que no hay diferencia entre la proporción de hombres y mujeres propietarios de vehículos

PREGUNTAS

- **¿Cómo y para que se emplean las pruebas no paramétricas?**
- **¿Cómo se elabora un análisis de datos con Chi²?**

Quinta unidad

……

ANALISIS COMBINADO

5.1. ANALISIS DE VARIANZA

Objetivo: El alumno al terminar el aprendizaje del presente capítulo sabrá como realizar el análisis de varianza.

VARIANZAS POBLACIONALES

Cuando se contrasta la hipótesis de igualdad de medias de dos poblaciones o cuando se realiza un análisis de la varianza (ANOVA) es fundamental decidir si puede aceptarse que las muestras independientes provienen de poblaciones con la misma varianza. Este problema se resuelve a partir del análisis exploratorio que proporciona los diagramas de caja y el estadístico del contraste de Levene.

Si la altura de las 'cajas' y los 'bigotes' correspondientes a los diagramas de caja de cada una de las muestras son aproximadamente iguales, se tiene un indicio de que posiblemente las muestras provienen de poblaciones con igual varianza.

Como complemento numérico al gráfico se realiza la prueba de Levene que calcula un estadístico que mide la diferencia entre las varianzas y la probabilidad de haberla obtenido al azar bajo el supuesto de que las varianzas poblacionales de los grupos sean iguales. Las hipótesis del contraste son:

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_n^2 \neq \sigma^2$$

La secuencia es:

Analizar

Estadísticos Descriptivos

Explorar

En el cuadro de diálogo se indica la variable de interés *'Dependiente'* y la variable que define los grupos *'Factores'*. En *Gráficos* se debe activar la opción *Estimación de potencia*.

El contraste de Levene se realiza por defecto cuando se contrasta la diferencia de dos o más medias.

EJEMPLO

Ejemplo 1.

Para la variable cuantitativa contrastar si existe diferencia significativa entre las varianzas según la clasificación de una variable cualitativa que tenga dos a más indicadores.

.

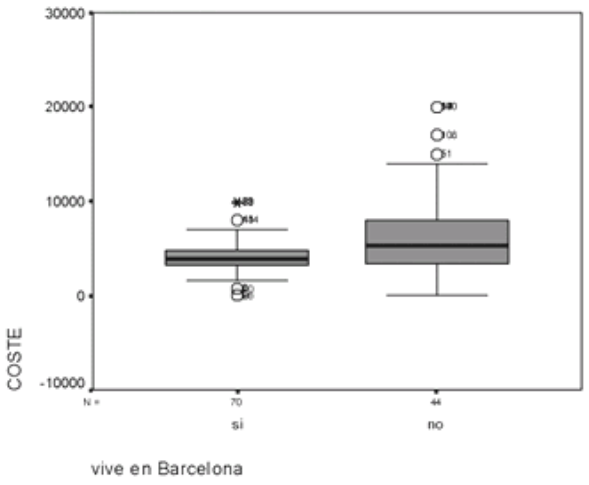
En el ejemplo 3 del epígrafe Diferencia de medidas poblacionales se trataba de verificar si existía una diferencia significativa entre el coste esperado en transporte de los alumnos que viven en Barcelona y el de los que viven fuera. En este caso es fundamental probar si las varianzas de ambos grupos pueden considerarse o no iguales, ya que de este supuesto depende que se deba escoger uno u otro de los dos estadísticos de prueba que aparecen en el cuadro de resultados del contraste.

Los resultados que se obtuvieron fueron los siguientes:

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
COSTE	Se han asumido varianzas iguales	37,671	,000	-4,459	112	,000
	No se han asumido varianzas iguales			-3,750	50,997	,000

Como puede verse, bajo la hipótesis nula de varianzas iguales el estadístico de Levene (F) toma el valor 37,671. Este valor es suficientemente grande como para rechazar la hipótesis nula para cualquier nivel de significación. Si se observan los correpondientes diagramas de caja:



Se ve claramente que la variabilidad del coste en el grupo de los residentes en Barcelona es menor que en el de los no residentes.

Ejemplo 2.

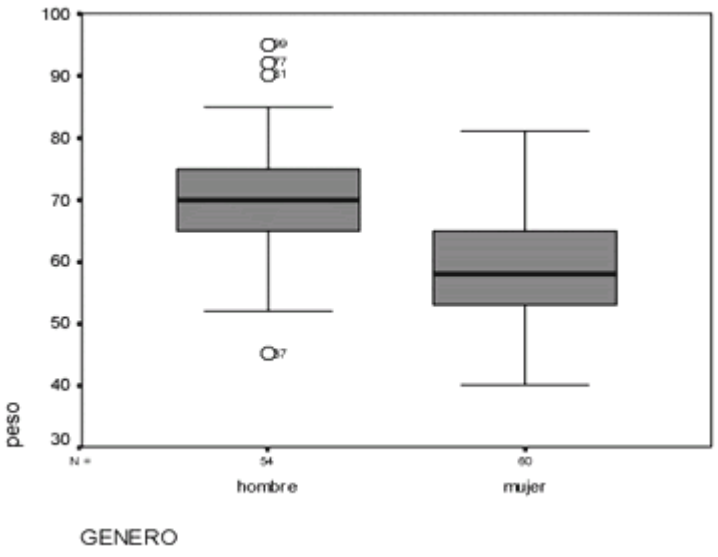
Se quiere comprobar si la distribución del Peso es o no más homogénea (presenta menor varianza) en el grupo de las mujeres que en el de los hombres.

Las hipótesis que se quieren contrastar son: $H_0{:}\sigma_H^2 = \sigma_M^2$ y $H_1{:}\sigma_H^2 > \sigma_M^2$.

La secuencia es *Analizar* > *Estadísticos Descriptivos* > *Explorar.*

Una vez seleccionadas las variables (Dependiente: Peso y Factor: Género) con el botón *Gráficos* se abre el cuadro de diálogo correspondiente en el cual se activa la opción *Estimación de potencia*.

En el cuadro *Descriptivos* que aparece en el visor de resultados se observa que la varianza correpondiente al grupo de mujeres es 76,368 y la de los hombres es 97,613. El contraste de Levene permite decidir si esta diferencia puede considerarse significativa o bien es debida únicamente al azar del muestreo.



El diagrama de caja (Box-Plot) pone de manifiesto que el recorrido del 50% de las observaciones centrales de la variable peso en el grupo de mujeres es algo mayor que en el grupo de hombres; pero, por otra parte, en este segundo grupo hay 4 casos outliers o extremos que si se tienen en cuenta determinan un recorrido total de la distribución mayor en este grupo. En consecuencia, esta primera aproximación gráfica, no permite llegar a ninguna conclusión.

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
peso	Basándose en la media	,214	1	112	,644
	Basándose en la mediana.	,243	1	112	,623
	Basándose en la mediana y con gl corregido	,243	1	108,912	,623
	Basándose en la media recortada	,228	1	112	,634

El estadístico F de la prueba de Levene (basándose en la media como valor central) es igual a 0,214, que a una cola presenta un nivel de significación crítico igual a 0,644. ésto significa que no puede rechazarse la hipótesis nula para los niveles de significación habituales y, por lo tanto, concluimos que la diferencia de varianzas muestrales no es significativa.

ANÁLISIS DE LA VARIANZA CON UN FACTOR (ANOVA)

El análisis de la varianza permite contrastar la hipótesis nula de que las medias de K poblaciones (K >2) son iguales, frente a la hipótesis alternativa de que por lo menos una de las poblaciones difiere de las demás en cuanto a su valor esperado. Este contraste es fundamental en el análisis de resultados experimentales, en los que interesa comparar los resultados de K 'tratamientos' o 'factores' con respecto a la variable dependiente o de interés.

$$H_0\colon\ \mu_1 = \mu_2 = \ldots = \mu_K = \mu$$
$$H_1\colon\ \exists \mu_j \neq \mu \quad j = 1, 2, \ldots, K$$

El Anova requiere el cumplimiento los siguientes supuestos:

- Las poblaciones (distribuciones de probabilidad de la variable dependiente correspondiente a cada factor) son normales.
- Las K muestras sobre las que se aplican los tratamientos son independientes.
- Las poblaciones tienen todas igual varianza (homoscedasticidad).

El ANOVA se basa en la descomposición de la variación total de los datos con respecto a la media global (SCT), que bajo el supuesto de que H0 es cierta es una estimación de σ^2 obtenida a partir de toda la información muestral, en dos partes:

- Variación dentro de las muestras (SCD) o Intra-grupos, cuantifica la dispersión de los valores de cada muestra con respecto a sus correspondientes medias.
- Variación entre muestras (SCE) o Inter-grupos, cuantifica la dispersión de las medias de las muestras con respecto a la media global.

Las expresiones para el cálculo de los elementos que intervienen en el Anova son las siguientes:

Media Global: $\overline{X} = \frac{\sum_{j=1}^{K} \sum_{i=1}^{nj} x_{ij}}{n}$.

Variación Total: $\mathrm{SCT} = \sum_{j=1}^{K} \sum_{i=1}^{nj} (x_{ij} - \overline{X})^2$

Variación Intra-grupos: $\mathrm{SCD} = \sum_{j=1}^{K} \sum_{i=1}^{nj} (x_{ij} - \overline{X_j})^2$

Variación Inter-grupos: $\mathrm{SCE} = \sum_{j=1}^{K} (\overline{X_j} - \overline{X})^2 n_j$

Siendo *xij* el i-ésimo valor de la muestra j-ésima; *nj* el tamaño de dicha muestra y $\overline{X_j}$ su media.

Cuando la hipótesis nula es cierta SCE/K-1 y SCD/n-K son dos estimadores insesgados de la varianza poblacional y el cociente entre ambos se distribuye según una F de Snedecor con K-1 grados de libertad en el numerador y N-K grados de libertad en el denominador. Por lo tanto, si H0 es cierta es de esperar que el cociente entre ambas estimaciones será aproximadamente igual a 1, de forma que se rechazará H0 si dicho cociente difiere significativamente de 1.

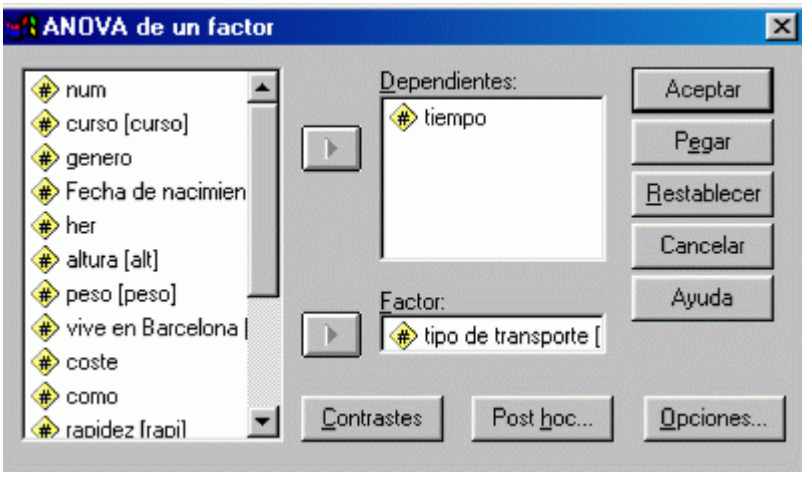
La secuencia para realizar un ANOVA es:

Analizar

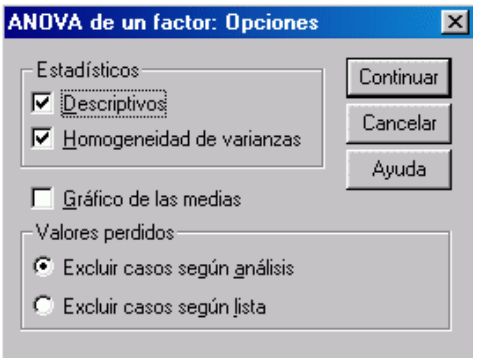
Comparar medias

ANOVA de un factor

Se abre el siguiente cuadro de diálogo:



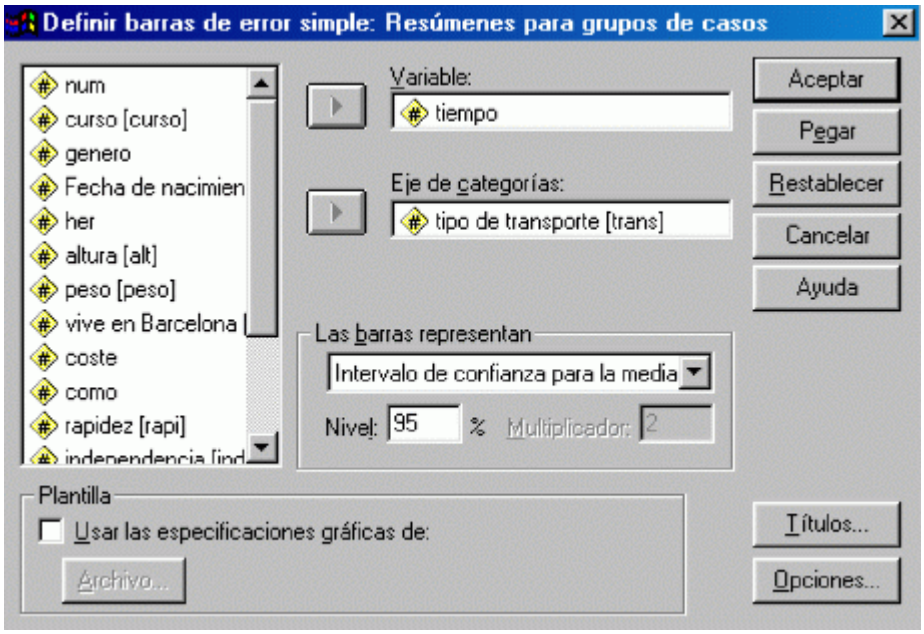
Se selecciona la variable que se considera *Dependiente* y la variable *Factor* y con el botón *Opciones* se activan *Estadísticos Descriptivos* y *Homogeneidad de varianzas.*



Al aceptar en el visor de resultados aparecen los siguientes cuadros:

- *Descriptivos.* Recoge la media, la desviación típica, el intervalo de confianza del 95% (por defecto) para la media correspondientes a la variable dependiente para cada uno de los grupos definidos por el factor.
- *Prueba de homogeneidad de varianzas*. Contiene el valor del estadístico de Levene del contraste de la hipótesis de homoscedasticidad con el nivel de significación crítico.
- *ANOVA*. Contiene las sumas de cuadrados inter-grupos, intra-grupos y total, sus correspondientes grados de libertad y el valor del estadístico de prueba F junto con el nivel de significación crítico.

Como complemento gráfico de este análisis, para obtener una primera aproximación acerca de si es razonable o no la hipótesis nula, se selecciona *Gráficos* > *Barras de error* y se activa la opción Simple. Con el botón *Definir* se abre el siguiente cuadro de diálogo:

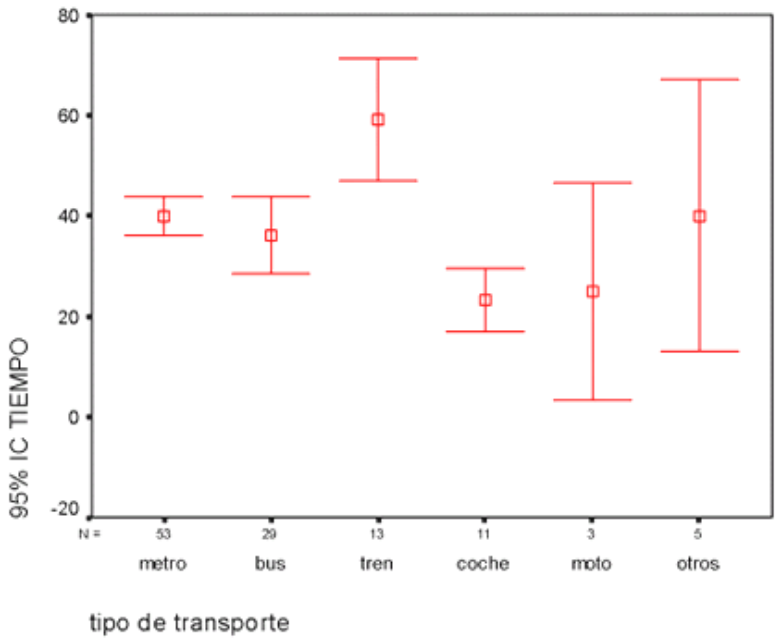


Se selecciona en *Variable* la variable dependiente del ANOVA y en el *Eje de categorías* la variable factor. El intervalo de confianza para la media se calcula por defecto al 95% de confianza. Al aceptar aparece en el visor de resultados los puntos que respresentan a la media de cada grupo junto con los límites del correspondiente intervalo de confianza para la media poblacional. Si los puntos que representan las medias están desigualmente distribuidos en el gráfico se tiene un indicio de que a nivel poblacional no puede sostenerse la hipótesis de igualdad de medias; es decir, por lo menos uno de los niveles del factor influye significativamente sobre la variable dependiente.

EJEMPLOS

Un ejemplo con transporte, razonar si puede aceptarse que el tipo de transporte utilizado, Trans, influye sobre la variable tiempo.

Con la opción de menú *Gráficos* > *Barras de error* > *Simple* y con el botón *Definir* se selecciona como *Variable* Tiempo y en *Eje de categorías* la variable Trans; al aceptar se obtiene la siguiente representación gráfica:



Como puede observarse, los puntos que representan a las medias de cada grupo aparecen dispersos a diferentes niveles; sobre todo la media del grupo definido por el factor Tren. El intervalo de confianza para la media correspondiente al grupo definido por el factor Metro está contenido dentro del intervalo correspondiente al grupo definido por el factor Bus, así como, el intervalo correspondiente al factor Coche está contenido dentro de los intervalos correspondientes definidos por los factores Metro y Otros. El gráfico, por tanto, parece sugerir no una única población sino tres poblaciones con distintas medias.

Para realizar el análisis de la varianza propiamente dicho la secuencia es *Analizar* > *Comparar medias* > *ANOVA de un factor*. En el cuadro de diálogo se selecciona Tiempo como variable *Dependiente* y Trans como *Factor*. Para contrastar la hipótesis de igualdad de varianzas se abre con el botón correspondiente el cuadro de diálogo *ANOVA de un factor*: *Opciones* y se activa *Homogeneidad de varianzas*. Si se desea un análisis descriptivo del comportamiento de la variable dependiente dentro de cada grupo se activa también la opción *Descriptivos*. Al aceptar se obtienen los siguientes cuadros de resultados:

Descriptivos

TIEMPO

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%	
					Límite inferior	Límite superior
metro	53	39,94	14,20	1,95	36,03	43,86
bus	29	36,17	20,18	3,75	28,50	43,85
tren	13	59,15	20,33	5,64	46,87	71,44
coche	11	23,18	9,56	2,88	16,76	29,60
moto	3	25,00	8,66	5,00	3,49	46,51
otros	5	40,00	21,79	9,75	12,94	67,06
Total	114	39,17	18,51	1,73	35,73	42,60

Este cuadro contiene un análisis descriptivo de la variable dependiente por grupos, así como, los límites superior e inferior para la media de cada grupo al 95% de confianza.

Prueba de homogeneidad de varianzas

TIEMPO

Estadístico de Levene	gl1	gl2	Sig.
1,514	5	108	,191

El estadístico de Levene toma un valor lo suficientemente pequeño para no rechazar la hipótesis de homocesdaticidad a los niveles de significación habituales.

ANOVA

TIEMPO

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	8901,537	5	1780,307	6,450	,000
Intra-grupos	29810,297	108	276,021		
Total	38711,833	113			

En el cuadro de resultados del ANOVA, el valor del estadístico de prueba, F=6,450, es significativamente distinto de 1 para cualquier nivel de significación y, por lo tanto, se rechaza la hipótesis nula de igualdad de medias y queda confirmada la primera impresión proporcionada por el gráfico de barras de error.

PREGUNTAS

- **¿Cómo se realiza un análisis de varianza ?**
- **¿Qué criterios se deben tener en cuenta para exponer realizar un análisis de varianza?**

5.2. REGRESION LINEAL

Objetivo: El alumno al terminar el aprendizaje del presente capítulo sabrá como realizar el análisis de regresión lineal.

MODELO DE REGRESIÓN LINEAL SIMPLE

En algunos casos la naturaleza de las variables permite suponer que existe relación de dependencia entre ellas, es decir, que los valores de una variable Y (variable dependiente o endógena) dependen o están influidos por los valores de otra variable, X (variable independiente o exógena). En el caso en que pueda suponerse una relación lineal de dependencia, ésta podrá sintetizarse mediante un modelo de regresión.

A partir del diagrama de dispersión y de los resultados obtenidos en el análisis de correlación puede decidirse si está relación es de tipo lineal. En este caso, los puntos del diagrama de dispersión aparecen tanto más próximos a una línea recta ajustada a la nube de puntos cuanto más intenso es el grado de asociación. Por otra parte, según sea el sentido de la asociación dicha línea tendrá pendiente positiva si el coeficiente de correlación simple, r, es positivo y negativa en caso contrario.

El punto de partida del modelo de regresión lineal simple (MRLS) es que la relación entre ambas variables no es de tipo determinista, sino estocástico; de forma que para cada valor de X existe una distribución de probabilidad de Y, siendo la relación tal que los valores esperados de las distribuciones de probabilidad de Y asociadas a cada uno de los valores de X están situados sobre una línea recta, llamada recta de regresión poblacional, que se expresa como: $E(Y_i) = \alpha + \beta X_i$.

ESTIMACIÓN

Para estimar la línea de regresión poblacional a partir de la nube de puntos se utiliza el método de los mínimos cuadrados ordinarios (MCO), que considera como recta que mejor se ajusta a la que minimiza la suma de los cuadrados de los resíduos.

Si la recta de mejor ajuste es $\widehat{Y}_i = a + bX_i$, los errores o resíduos se definen como: $e_i = Y_i - \widehat{Y}_i$; y los estimadores por MCO de la ordenada en el origen, α, y de la pendiente, β, son:

$$a = \overline{Y} - b\overline{X} \qquad b = \frac{S_{XY}}{S_X^2}$$

Para evaluar la bondad del ajuste se calcula el coeficiente de determinación R^2 y, para medir la dispersión de los puntos alrededor de la recta estimada, el error típico de la estimación Su. Estas medidas se definen como:

$$S_u = \sqrt{\frac{\sum e_i^2}{n-2}} \qquad R^2 = \frac{SCR}{SCT}$$

Donde SCT o suma total de cuadrados es la variación total de Y en la muestra y SCR o suma de cuadrados de la regresión es la parte de la variación total explicada por la recta ajustada. Por lo tanto, R^2 indica la proporción de variación total explicada mediante la relación lineal entre X e Y, y toma valores entre 0 y 1. Un valor de R^2 próximo a 1 indica que la recta ajustada es un buen modelo para explicar el comportamiento de la variable Y, y por lo tanto existe relación lineal entre X e Y. Por el contrario, un valor próximo a 0 indica que la recta ajustada no explica la variación observada en Y.

Para establecer el intervalo de confianza para la pendiente de la recta de regresión, β, y contrastar si el valor de este parámetro es o no significativamente diferente a cero es necesario calcular el error típico de b que se define como:

$$S_b = S_u \sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

El estadístico de prueba del contraste es $t = \frac{b}{S_b}$ que presenta una distribución de probabilidad t de Student con n-2 grados de libertad.

Para la obtención de la recta de regresión la secuencia es:

Analizar

Regresión

Lineal

Se abre el cuadro de diálogo *Regresión lineal* donde se seleccionan las variables *Dependiente* e *Independientes.*

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,018	,083	,212	,832
	Tau-c de Kendall	,017	,079	,212	,832
	Gamma	,020	,096	,212	,832
	Correlación de Spearman	,028	,106	,282	,778[c]
Intervalo por	R de Pearson	,111	,115	1,127	,262[c]
N de casos válidos		104			

a. No asumiendo la hipótesis nula.

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

c. Basada en la aproximación normal.

- La opción *Método* permite elegir el método de estimación. Si se trata de una regresión lineal simple (con una sola variable independiente) se conserva la definida por defecto (*Introducir*) siendo el resto de opciones para modelos con más de una variable explicativa.

- Cuando se desee realizar un ajuste lineal basado únicamente en los casos que pertenecen a un subgrupo determinado por un valor o conjunto de valores de otra variable, ésta se deberá indicar en *Variable de selección* del cuadro de diálogo *Regresión linea*l e introducir la Regla o condición que debe verificar un caso para ser incluido en el análisis.

- Opcionalmente se puede seleccionar la variable que recoge las etiquetas de los casos indicándola en *Etiquetas de caso.*

- El botón *MCP* hace referencia a la estimación por mínimos cuadrados ponderados.

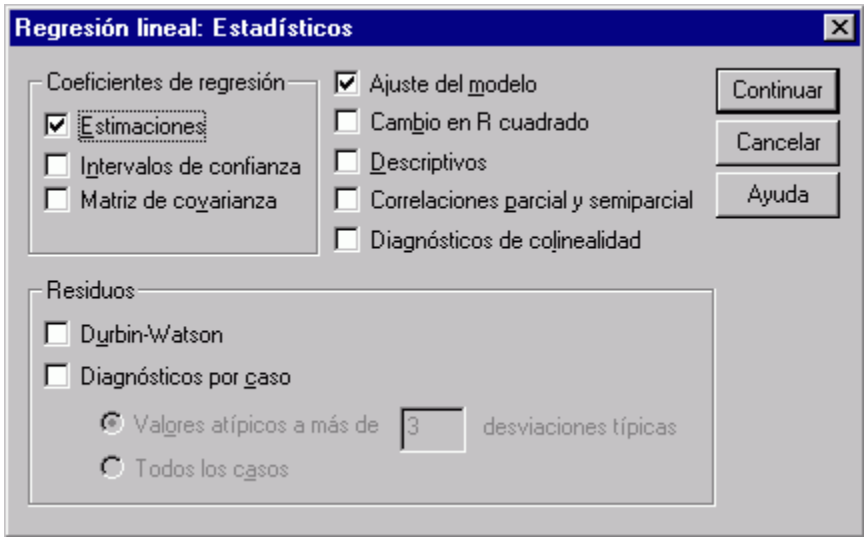
Este cuadro de diálogo además permite ampliar el análisis de regresión activando las opciones incluidas en Estadísticos, Gráficos y Opciones.

ESTADÍSTICOS

El botón *Estadísticos* abre el cuadro de diálogo *Regresión lineal*: *Estadísticos* que por defecto tiene activadas las opciones *Estimaciones* y *Ajuste del modelo*.

- La opción *Estimaciones* proporciona las estimaciones de los coeficientes de la recta ajustada $\widehat{Y}_i = a + bX_i$, por el método de los mínimos cuadrados ordinarios y sus correspondientes errores típicos, así como los coeficientes estandarizados (beta), los valores del estadístico t y el nivel de significación crítico.

- La opción *Ajuste del modelo* muestra en el resumen del modelo la bondad del ajuste o coefiente de determinación y en el cuadro ANOVA la descomposición de la suma total de cuadrados o información total observada.



Otras opciones que presenta este cuadro de diálogo son:

- *Intervalos de confianza* de los coeficientes de regresión que por defecto se calculan al 95\%.

- *Matriz de covarianzas* y de varianzas, y la matriz de correlaciones de los coeficientes del modelo que se analiza en el contexto de la regresión múltiple.

- *Cambio en R cuadrado*. Cuantifica la variación del coeficiente de determinación que se produce al añadir o eliminar alguna variable independiente en un modelo de regresión múltiple.

- *Descriptivos* incluye las medias y las desviaciones típicas de las variables seleccionadas y la matriz de correlaciones.

- *Diagnósticos por caso.* Esta opción presenta dos alternativas para el análisis de los residuos:

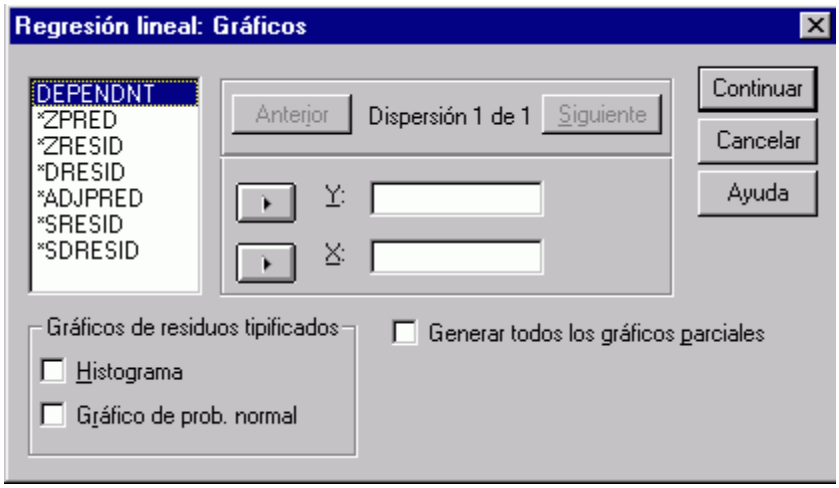
- la obtención de *Valores atípicos a más* de (por defecto 3) *desviaciones típicas*. Identifica aquellos casos para los cuales el valor estandarizado de los residuos difiere en (por defecto 3) o más desviaciones típicas de su media. Para estos casos también presenta el valor observado, el valor pronosticado y el residuo sin estandarizar. Incluye un cuadro de estadísticos de los residuos con la media y desviación típica de los valores pronosticados y de los residuos, tipificados y no tipificados, diferenciando entre los casos incluidos y los excluidos del análisis. La identificación de casos atípicos es importante porque su presencia en la muestra puede distorsionar los resultados de la regresión.

- la obtención de predicciones de Y para *Todos los casos*. Genera las predicciones de Y y sus correspondientes residuos para todos los casos.

El resto de opciones hacen referencia al modelo de regresión lineal múltiple.

GRÁFICOS

El botón *Gráficos* abre el cuadro de diálogo *Regresión Lineal: Gráficos.*



Este cuadro de diálogo permite seleccionar los gráficos a incluir en los resultados.

El recuadro superior presenta una serie de nuevas variables relacionadas con las predicciones y los residuos. éstas pueden ser seleccionadas para definir los ejes X e Y de los diagramas de dispersión que se quieren elaborar. Pulsando el botón *Siguiente* el programa va numerando los diagramas que incluirá en los resultados.

El recuadro *Gráficos de residuos tipificados* presenta dos opciones: *Histograma* que muestra un histograma de los residuos tipificados superponiéndole la distribución

normal y *Gráfico de prob. normal* que crea un gráfico P-P útil para comprobar la hipótesis de normalidad a partir de los residuos tipificados. La comprobación de esta hipótesis es fundamental para la correcta interpretación de las estimaciones por intervalo, tanto de los coeficientes de la recta como de las predicciones.

OPCIONES

El botón *Opciones* abre el cuadro de diálogo *Regresión Lineal: Opciones.*

Correlaciones

		EST	PESO
EST	Correlación de Pearson		
	Sig. (bilateral)		
	N		
PESO	Correlación de Pearson	,883**	
	Sig. (bilateral)	,000	
	N	114	

**. La correlación es significativa al nivel 0,01

Permite desactivar *Incluir constante en la ecuación* que elimina el término independiente y proporciona la recta de regresión que pasa por el origen de coordenadas. Por lo que se refiere a los *Valores perdidos*, además de las dos posibilidades *Excluir casos según lista*, activada por defecto, y *Excluir casos según pareja*, comentadas en el epígrafe 3.6, hay la posibilidad de *Reemplazar por la media*, opción que sustituye los valores missing por la media de la variable correspondiente.

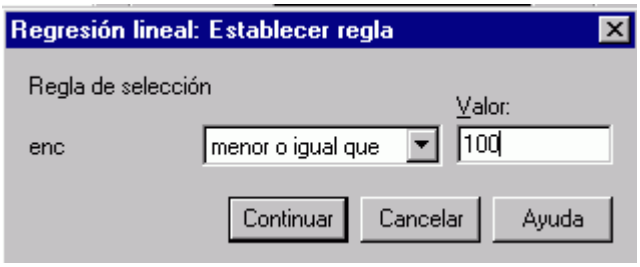
- El bloque *Intervalos de pronóstico* calcula intervalos de confianza para las predicciones de la Media y/o los *Individuos* para el nivel de confianza deseado (95% de confianza por defecto).
- Si se desea guardar estos resultados en un archivo nuevo se activa la opción *Estadísticos de los coeficientes* y se indica el nombre del archivo.

EJEMPLOS

Ejemplo 1.

Con las variables Peso y Est (estatura) estime el modelo de regresión lineal simple que explica el comportamiento del Peso (variable dependiente) en función de la Est (variable independiente). Realice la estimación con los 100 primeros casos.

Con la secuencia *Analizar* > *Regresión* > *Lineal* aparece el correspondiente cuadro de diálogo en el que se seleccionan la variable Peso como *Dependiente* y la variable Est como *Independiente*. En el recuadro Variable de selección se introduce la variable Enc (número de encuesta) y con el botón Regla se abre el cuadro de diálogo *Regresión Lineal: Establecer regla* donde se introduce la condición 'menor o igual que 100'.



Los resultados que se obtienen son:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error tip. de la estimación
	Número de encuesta <= 100 (Seleccionado)			
1	,883[a]	,780	,778	6,0638

a. Variables predictoras: (Constante), EST

En el cuadro resumen del modelo se observa que: r=0,883, R^2=0,78 (obsérvese que R^2 es igual a r al cuadrado) y Su=6,0638. El coeficiente de determinación indica que el 78% de la variación total del peso en la muestra queda explicada por el modelo estimado y, por lo tanto, el modelo proporciona un buen ajuste.

Coeficientes[a, b]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error tip.	Beta		
1	(Constante)	-132,783	10,613		-12,511	,000
	EST	1,148	,062	,883	18,662	,000

a. Variable dependiente: PESO

b. Seleccionando sólo los casos para los que Número de encuesta <= 100

El cuadro *Coeficientes* presenta los siguientes resultados:

- Modelo estimado: $\widehat{Peso}$=-132,783 + 1,148*Est.*

- Errores típicos (errores estándar) de las estimaciones de los parámetros α y β: Sa=10,613 y Sb=0,062.

- Coeficientes beta, que se obtienen estimando la regresión a partir de las observaciones estandarizadas. En la regresión simple este coeficiente coincide con el coeficiente de correlación lineal simple, r.

- Estadísticos t de los contrastes de significación de las estimaciones y sus correspondientes niveles de significación críticos: $t_a = \frac{a}{S_a}$ =-12,511 y $t_b = \frac{b}{S_b}$ =18,662. En este caso las estimaciones son significativamente distintas decero para cualquier nivel de significación.

Ejemplo 2.

Compruebe si existen valores extremos y analice el comportamiento de los residuos del modelo de regresión lineal estimado en el apartado anterior.

Con la secuencia *Analizar* > *Regresión* > *Lineal* aparece el correspondiente cuadro de diálogo en el que se mantienen seleccionadas la variable Peso como *Dependiente* y la variable Est como *Independiente.* Con el botón *Estadísticos* se accede al cuadro de diálogo que presenta las opciones correspondientes al diagnóstico de residuos. Se activa *Diagnóstico por caso* y *Valores atípicos a más de 2 desviaciones típicas.*

Diagnósticos por caso[a]

Número de caso	Residuo tip.	PESO	Valor pronosticado	Residual
63	2,046	100,00	87,5962	12,4038

a. Variable dependiente: PESO

Estadísticos sobre los residuos[a,b]

	Número de encuesta <= 100 (Seleccionado)					Número de encuesta > 100 (No seleccionado)				
	Mínimo	Máximo	Media	Desviación típ.	N	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	37,0927	88,7440	64,95	11,373	100	43,9796	81,8572	61,28	10,977	14
Residual	-10,823	12,4038	,0000	6,0331	100	-6,5269	10,0204	,6499	4,8188	14
Valor pronosticado tip.	-2,449	2,092	,000	1,000	100	-1,844	1,487	-,323	,965	14
Residuo tip.	-1,785	2,046	,000	,995	100	-1,076	1,653	,107	,795	14

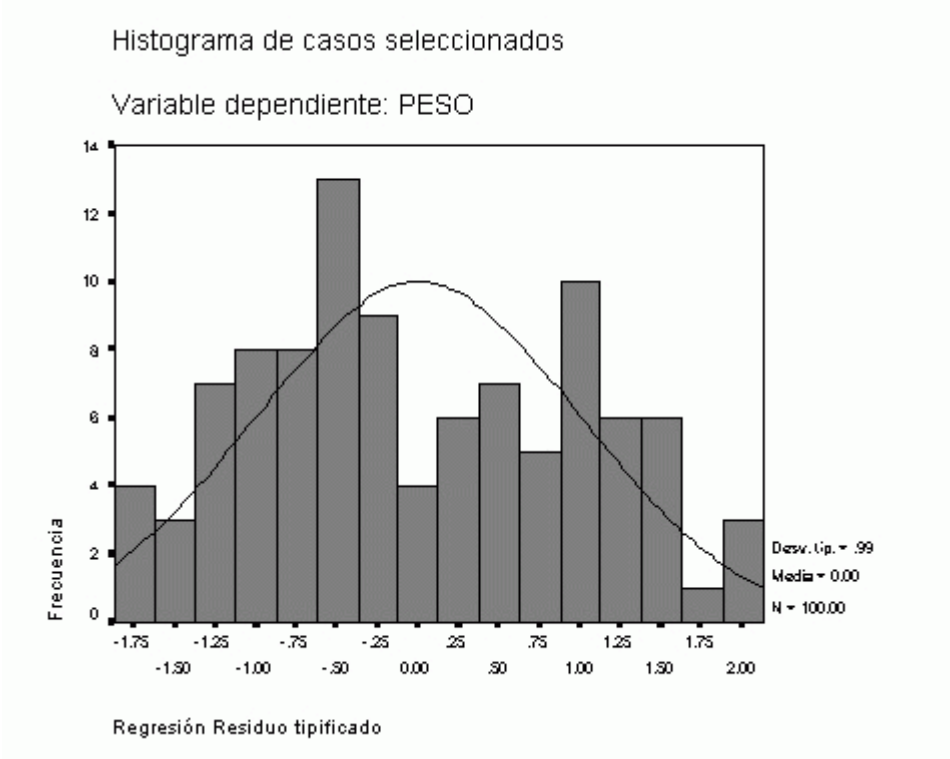
a. Variable dependiente: PESO

b. Casos combinados

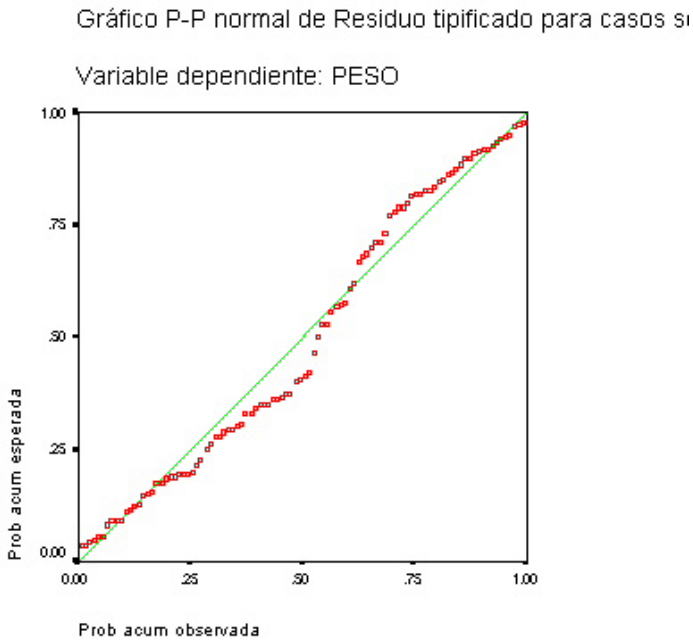
- Se observa que únicamente un caso presenta un resíduo estandarizado, igual a 2,046, superior a 2 veces la desviación estándar. Esto nos indica que no existe ningún caso atípico.

- En el cuadro *Estadísticos sobre los residuos* se comprueba que efectivamente no hay valores atípicos ya que los valores máximo y mínimo de los residuos tipificados son inferiores a 3 en valor absoluto.

Con el botón *Gráficos* se abre el cuadro de diálogo donde se deben activar las opciones correspondientes a los *Gráficos de residuos tipificados.*

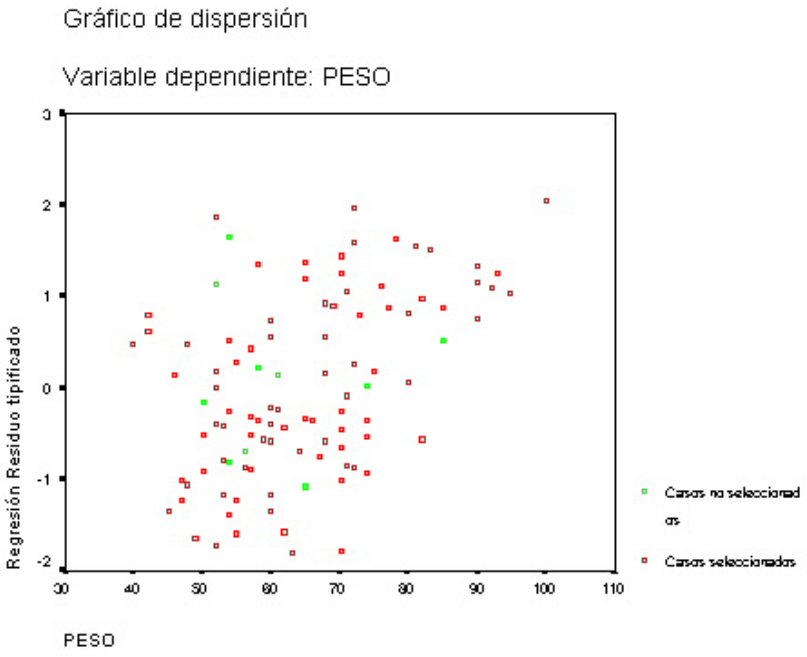


El histograma de los residuos permite comprobar gráficamente la hipótesis de normalidad; aspecto que deberá tenerse en cuenta para la interpretación de los resultados de la inferencia estadística. En este caso vemos que la distribución es campaniforme pero presenta una laguna en el centro que puede ser, en parte, consecuencia de los intervalos definidos.



El diagrama P-P compara la frecuencia acumulada por los residuos tipificados con la probabilidad esperada bajo la hipótesis de normalidad. Se observa que estas diferencias podrían ser significativas en alguna zona del gráfico; lo cual, de ser cierto, pondría en duda la validez de la hipótesis de normalidad de los residuos. No obstante, el criterio para decidir si se puede rechazar la hipótesis de normalidad será el que proporcione alguno de los contrastes de normalidad.

Además, en el mismo cuadro de diálogo se puede pedir que elabore los diagramas de dispersión de, por ejemplo, los residuos estandarizados en función de la variable dependiente (ZRESID y DEPENDNT).



En el gráfico vemos que no existe ningún patrón de comportamiento de los residuos respecto a Y. Por lo tanto, podemos mantener que estas variables aleatorias están incorrelacionadas.

Ejemplo 3.

Obtenga las predicciones y los residuos correspondientes a los 100 casos incluidos en la estimación y a los 14 casos excluidos.

Con la secuencia *Analizar* > *Regresión* > *Lineal* aparece el correspondiente cuadro de diálogo en el que se mantienen seleccionadas la variable Peso como *Dependiente* y la variable Est como *Independiente*. Con el botón *Guardar* se abre el cuadro de diálogo donde se deben activar las opciones:

- *Valores Pronosticados* > *No tipificados* y *Tipificados* para obtener las predicciones a partir del modelo estimado.
- *Residuos* > *No tipificados* y *Tipificados* para obtener los residuos de todos los casos.
- *Intervalos de Pronóstico* > *Media* e *Individuos* para obtener los límites de los intervalos.

Los resultados de estas opciones quedan almacenados en el archivo de datos activo, y están disponibles para análisis posteriores. Por defecto los nombres de las variables que crea son: Pre_1 (predicciones no estandarizadas), Res_1 (residuos no estandarizados), Zpr_1, Zre_1(predicciones y residuos estandarizados, respectivamente), Sep_1 (error estándar de las predicciones), Imci_1, Unci_1 (Límite inferior y superior del intervalo de confianza para la predicción del valor esperado de Y), Lici_1, Uici_1 (Límite inferior y superior del intervalo de confianza para la predicción individual de Y).

Por ejemplo, para el caso 101, que presenta una estatura de 168 y un pesoigual a 56, los resultados son:

- Predicción del peso sin tipificar 60,04886 Kg. y tipificada -0,43094.

- Residuos, no tipificados y tipificados, -4,04886 y -0,66771, respectivamente.

- Error estándar de la predicción 0,66081.

- Intervalo de confianza para el valor esperado de Y para los individuos con estatura 168 (58,73751 ; 61,36022).

- Intervalo de confianza para el valor individual (47,94423 ; 72,15350).

Idénticamente, la predicción para el caso 102, que presenta una estatura de 180, es de 73,82255 kg., con un residuo igual a -3,82255 y un error estándar 0,77055. Los correspondientes intervalos de confianza para el valor esperado y para el valor individual son (72,29343 ; 75,35167) y (61,69239; 85,95271), respectivamente.

BIBLIOGRAFÍA

1. Eco, Umberto. Cómo se hace una tesis. Técnicas y procedimientos de estudio, investigación y escritura, Gedisa, Barcelona, 2001.
2. Cassany, Daniel. La cocina de la escritura, Anagrama, Barcelona, 1995.
3. Harvey, Gordon. Cómo se citan las fuentes, Nuer, Madrid, 2001.
4. Hoffman, Eric. Guidebook for Publishing Philosophy, Philosophy Documentation Center, Bowling Green, OH, 1997.
5. Izuzquiza, Ignacio. Guía para el estudio de la filosofía. Referencias y métodos, Anthropos, Barcelona, 1989.
6. List, Charles J. y Plum, Stephen H. Library Research Guide to Philosophy, Pierian Press, Ann Arbor, MI, 1990.
7. Martinich, Aloysius P. Philosophical Writing. An Introducction, Prentice Hall, Englewood Cliffs, NJ, 1990.
8. Seech, Zachary. Writting Philosophy Papers, Wadsworth, CA, 1997.
9. Turabian, Kate L. A Manual for Writers of Term Papers, Theses and Dissertations, The University of Chicago Press, Chicago, 1987.
10. Watson, Richard A. Writting Philosophy. A Guide to Professional Writing and Publishing, Southern Illinois University Press, Carbondale, IL, 1992.
11. Chacón, Albam: Los trabajos finales de graduación, Guías No.1 y No.2, Escuela de Administración Educativa, Universidad de Costa Rica, 1986.
12. Chacón, Albam: Los trabajos finales de graduación su elaboración y presentación en las Ciencias Sociales, Editorial Universidad Estatal a Distancia, ISBN 9977-64-323-7, San José, Costa Rica, 1991.
13. Cerdas, Manuel: HTML.tab: Actualización y despliegue de tablas tridimensionales en WWW, Propuesta de Trabajo Final de Graduación, Licenciatura en Ciencias de la Computación e Informática, Escuela de Ciencias de la Computación e Informática [ECCI], UCR-1997. Disponible en Internet en: http://www.di-mare.com/adolfo/cursos/mcerdas.htm
14. Di Mare, Adolfo: Guía para tipografiar artículos en Internet, Revista Acta Académica, Universidad Autónoma de Centro América, Número 21, pp [26-37], ISSN 1017-7507, Noviembre 1997.
15. Di Mare, Adolfo: Temas de tesis, Noviembre 1997. Disponible en Internet en: http://www.di-mare.com/adolfo/cursos/tesis.htm

16. Levine, S. Joseph: Como Escribir y Presentar su Tesis o Disertación, Traducido por Ernesto Restaino Instituto Nacional de Investigación Agropecuaria (INIA) Colonia, Uruguay. Disponible en Internet en: http://www.learnerassociates.net/dissthes/guidesp.htm
17. Univesridad de Costa Rica, Sistema de Estudios de Posgrado: Reglamento de Tesis del Sistema de Estudios de Posgrado, [Aprobado en sesión 2469-07, 05-04-78, publicado en la Gaceta Universitaria 9, Año II, 12-05-78]. Disponible en Internet en: http://cu.ucr.ac.cr/normativ/tesis_del_sep.pdf
18. Univesridad de Costa Rica: Reglamento de Trabajos Finales de Graduación, [Aprobado en sesión No. 2713-17, 4-8-80, publicado como anexo 1 del acta respectiva]. Disponible en Internet en: http://cu.ucr.ac.cr/normativ/trabajos_finales_graduacion.pdf
19. Zorrilla Arena, Santiago & Torres Xammar, Miguel: Guía para elaborar la tesis, McGraw-Hill Interamericana de México, ISBN 970-10-0139-7, 1992.
20. Zubizarreta, Armando: La Aventura del Trabajo Intelectual, 1987.
21. Bayarre, H. y cols. (2004) Metodología de la investigación en la APS,
22. Comisión Nacional de Grados Científicos. (2005) Normas para la redacción y presentación de las tesis de Doctor en Ciencias de determinada especialidad. En Normas para la obtención de Grados científicos. República de Cuba. Pág. 47 -52.
23. Eco, Umberto.(1991) “Cómo se hace una tesis”de Ed. Gedisa España, p.188.
24. Eco, Umberto (1991) “Cómo se hace una tesis”de Ed. Gedisa España, p. 201
25. Eco, Umberto (1991) “Cómo se hace una tesis”de Ed. Gedisa España, pág199.
26. Hernández, E. Palomera, A; de Santos, F. (2003) Intervención psicológica en las enfermedades cardiovasculares. Editora Universidad de Guadalajara, Jalisco, México.
27. Hernández, E.; Grau, J. y cols. (2005). Psicología de la Salud. Fundamentos y aplicaciones. Editorial La Noche. Guadalajara, Jalisco, México.
28. Referencias bibliográficas según el Estilo Vancouver. Biblioteca de la ENSAP.
29. Torres, M. (2005) Taller de Tesis II. Bibliografía básica. Material redactado para el Dossier de la Maestría en Salud Familiar y Comunitaria,

Jully Pahola Calderón Saldaña Ph. D.
Doctor of Philosophy In Public Health
Doctora en Salud Pública.
Maestra en Obstetricia con mención en Salud Reproductiva.
Diploma en Educación Médica.

Luis Alzamora de los Godos Urcia Ph. D.
Doctor of Philosophy In Public Health
Doctor en Salud Pública.
Maestro en Obstetricia con mención en Salud Reproductiva.
Diploma en Educación Médica.

El Libro de estadística cambia el paradigma de la estadística teórica y se concentra en convertirla en un sistema de aprendizaje práctico y útil, ya que el candidato a maestro no debe ser un matemático estadístico, sino un profesional de la educación que emplee la estadística de manera práctica y sencilla para trabajar su tesis, en este sentido se le facilitará los elementos teórico prácticos para utilizar un programa informático para llevar a cabo el tratamiento y análisis de información estadística. Se dirige a un conjunto muy amplio de estudiantes de maestría, tanto aquellos que se inicien en el aprendizaje de la Estadística como para los que ya tienen unos conocimientos previos sobre la materia y quieren aplicarlos con la ayuda de un programa ampliamente difundido en la actualidad como es el programa SPSS.

Se presupone que el participante de maestría que utiliza esta aplicación quiere introducirse en los conocimientos de la Estadística mediante la utilización de un programa informático para el tratamiento de datos, concretamente el programa SPSS, versión 11. Para el seguimiento del módulo no se requiere ningún conocimiento previo del funcionamiento de este programa, y muy pocos conocimientos de matemática básica. Este material ha sido concebido como un instrumento aplicado al aprendizaje de la Estadística, ya que permite ver cómo se aplican los conocimientos y se obtienen los resultados con las herramientas informáticas disponibles.

En cada uno de los apartados se consideran dos partes que permiten, en primer lugar, familiarizarse con el entorno del programa SPSS, y seguidamente se procede a explicar las técnicas de análisis de datos: se incluyen una explicación teórica con definiciones, expresiones y fórmulas que permite introducir o recordar al lector la teoría estadística que se está utilizando.

ISBN 978-0-557-97604-1
90000