

INDICE

INTRODUCCIÓN	7
DESCRIPCIÓN DE VARIABLES: MEDIDAS.	11
Clasificación de variables	14
La tabla de distribución de frecuencias	16
Estadísticos resúmenes de distribuciones	20
Medidas de tendencia central	20
Medidas de localización	21
Medidas de simetría	28
Medidas de apuntamiento	30
Los estimadores robustos centrales	32
Obtención de las distribuciones y sus medidas.	34
VISUALIZACIÓN DE VARIABLES: GRÁFICOS	41
Gráficos para variables discretas	44
Gráficos para variables continuas	53
TRANSFORMACIÓN DE VARIABLES	63
Instrucciones para transformar variables	72
COMPARACIÓN Y ASOCIACIÓN DE VARIABLES	77
Análisis de variables cualitativas	79
Análisis de variables cuantitativas.	88
Supuestos de la regresión	105
Relaciones no lineales	110
OTRAS TÉCNICAS EXPLORATORIAS	131
BIBLIOGRAFÍA	135

INTRODUCCIÓN

La Estadística es llamada a veces la Ciencia de los grandes números, pues entre sus múltiples operaciones se encuentra la función de explorar, resumir y encontrar pautas existentes en grandes cantidades de datos. Las características de millones de personas pueden sintetizarse diciendo que un 52% de la población son mujeres y un 48% son varones. Otra de las principales funciones de la Estadística es la de clasificar a conjuntos de seres u objetos. Es decir, mediante los procedimientos adecuados pueden agruparse a los individuos de modo que queden conformados en bloques homogéneos. La Estadística también ayuda a demostrar las conexiones causales existentes a través del estudio de las asociaciones de fenómenos o características y sirve en consecuencia para la realización de predicciones de la realidad. Finalmente, para cumplimentar estas funciones, no es preciso recoger el conjunto de datos posibles (población), sino que con una adecuada (generalmente aleatoria) selección de casos, se pueden realizar correctas inferencias sobre el conjunto. En definitiva, el análisis de datos estadístico persigue la descripción, comprensión, explicación o previsión a través del resumen o la relación de las informaciones obtenidas de una serie determinada de objetos.

En la realización de análisis estadísticos pueden distinguirse dos aproximaciones: la primera va a ser llamada análisis exploratorio de datos (AED) e incluye la organización, descripción y resumen de los datos. Mediante ella, a través de análisis numéricos y representaciones gráficas se destacan las características que presentan los datos a fin de detectar posibles errores en ellos y, sobre todo, descubrir cuáles son las pautas que los caracterizan.

La segunda aproximación se basa en la estimación, prueba de hipótesis y ajuste de modelos que den cuenta de la distribución obtenida de los datos. Es evidente que las investigaciones mejores y más desarrolladas incluyen necesariamente las operaciones de esta segunda etapa; pero tampoco es menos cierto que estas operaciones no pueden llevarse a cabo sin previamente haber realizado un adecuado análisis exploratorio de datos.

A partir del libro que Tukey publicó en 1977, titulado *Exploratory Data Analysis*, los estadísticos han prestado cada vez más atención a un examen atento de los datos previo a la realización de técnicas más complejas de confirmación de hipótesis. Este libro incorporó una nueva perspectiva al análisis estadístico porque, además de añadir nuevos índices para el análisis de las variables, se basaba en los siguientes criterios:

a) Una importancia central otorgada a la representación gráfica de las variables. Un modo simple de formarse una idea de cómo se distribuyen los datos es mediante su transformación en un gráfico porque de esa forma no sólo se perciben las puntuaciones centrales sino también la particular configuración del conjunto de individuos.

b) Un énfasis puesto en la calidad de la resistencia de los estadísticos, característica que consiste en la insensibilidad de éstos a la presencia de datos extremos o, incluso, errores altos de medición. Por ello, si la estadística clásica se centraba en medidas como la media aritmética y la varianza, el análisis exploratorio de datos prefiere la mediana y la amplitud intercuartiles, pues son más resistentes a casos extremos.

c) Una concepción inductiva de la estadística basada en la distinción entre el ajuste y el residuo. Lo primero es aquello que sigue una determinada pauta, lo segundo la parte del dato que se aleja del ajuste. El análisis exploratorio intenta buscar el mejor ajuste posible a los datos a través de distintas pruebas iterativas de modo que los residuos sean los mínimos posibles.

d) Una apertura a la transformación de las variables con el objeto de conseguir modelos más ajustados. Desde la eliminación de los casos extremos en las variables hasta la reconversión no lineal de éstas, el análisis exploratorio usa técnicas que transforman los datos para la consecución de ajustes más notables.

Este libro pretende ser una guía para estudiantes e investigadores que, sin tener amplios conocimientos matemáticos ni estadísticos, deseen conocer las herramientas gráficas y exploratorias. La exposición del contenido se hace desde un punto de vista integrador pues presenta y explica los índices y gráficos tanto de la escuela clásica de la Estadística como de la del análisis exploratorio de datos. Por otro lado, consciente de la importancia e inevitabilidad de la utilización de programas informáticos, se incorpora la explicación de cómo se realizan los gráficos y se obtienen los estadísticos. Para seguir el estilo de la colección donde se integra esta obra, se ha elegido el programa SPSS¹, que incorpora un programa básico de análisis, llamado *Examinar*, con un importante núcleo de técnicas exploratorias.

El primer capítulo comienza con los rudimentos del análisis: las variables, consideradas aisladamente una por una, el estudio de su distribución y las medidas que las caracterizan. Seguidamente, se presenta un repertorio de los gráficos para la representación de las variables. El tercer capítulo pasa revista a las transformaciones que se pueden aplicar a los datos para obtener mejores ajustes. En el último capítulo, se desarrollan los procedimientos que se utilizan para el análisis conjunto de variables: la comparación y la asociación.

Para mostrar ejemplos de los procedimientos explicados, se han utilizado tres fuentes distintas. En primer lugar, se han expuesto datos sencillos e imaginarios para que el lector capte inmediatamente las operaciones estadísticas practicadas. Hay una segunda fuente, que consiste en datos socioeconómicos de países europeos extraídos del Atlas del Banco Mundial. Por último, cuando hay ejemplos de datos individuales de cuestionario, se utiliza la última encuesta que realizó el Cires (Centro de Investigación de la Realidad Social) en junio de 1996 bajo la dirección de Juan Díez Nicolás con una muestra de 1200 españoles mayores de 18 años.

¹ La versión que se ha utilizado es la 8.01 en castellano.

DESCRIPCIÓN DE VARIABLES: MEDIDAS.

Cada uno de los entes de los que se dispone información recibe el nombre de *unidad* o caso. A cada unidad se puede atribuir un determinado *valor*. El conjunto de valores que de forma exhaustiva y mutuamente excluyente se puede atribuir a todas las unidades que se analizan recibe el nombre de *variable*. Expresado con distintas palabras, una variable es cualquier característica susceptible de adquirir distintas modalidades o valores. De esta forma, una serie de lápices puede clasificarse según su color (variable) en negro, rojo, verde, azul, blanco, ... (valores). También las personas pueden distinguirse según sean mujeres u hombres, tengan una edad u otra, sostengan una opinión o su contraria.

La información bruta de todo análisis estadístico es la matriz de datos. Esta consiste en una forma de ordenación de la información basada en dos entradas o dimensiones. En una de ellas, generalmente la horizontal, se disponen las unidades; en la otra, generalmente la vertical, las variables. En la intersección de ambas se expresa el valor que una específica unidad tiene en una variable. Sea, por ejemplo, una familia con cuatro personas: la madre de 34 años, el padre con 35 y dos hijos varones con 5 y 6 años respectivamente. La matriz de datos que describiría cada una de las unidades en el conjunto tendría cuatro líneas y dos columnas. Sus dimensiones serían de 4x2.

Tabla 1.- Matriz de datos

UNIDADES	VARIABLES	
	SEXO	EDAD
Madre	Mujer	34
Padre	Varón	35
Hijo mayor	Varón	6
Hijo menor	Varón	5

La matriz de datos de la Tabla 1 se compone de los ocho valores encerrados en el rectángulo: los cuatro, en realidad sólo dos distintos, de la primera columna corresponden a la variable sexo; los cuatro de la segunda columna pertenecen a la variable edad.



	sexo	edad	var
1	Mujer	34	
2	Hombre	35	
3	Hombre	6	
4	Hombre	5	
5			

Figura 1.- Matriz de datos en SPSS.

Desde el punto de vista informático, una matriz de datos equivale a un fichero con características y extensión diferencial (.sav, por omisión en el programa SPSS). En la figura 1, se muestra el fichero *familia* que contiene cuatro filas, reconocidas cada una de ellas con el número de orden y dos columnas, que representan sendas variables, denominadas respectivamente cada una de ellas sexo y edad².

Clasificación de variables

Las variables pueden ser clasificadas de distintos modos según las características que tengan los valores. En el primer ejemplo de matriz de datos aparecieron dos variables, que aparentemente son diferentes. La primera tiene valores de naturaleza *cualitativa*: "Varón" y "Mujer". La segunda presenta valores de tipo *cuantitativo*: "5", "6", "34" y "35". No cabe la menor duda de que el tratamiento que se puede aplicar a una y otra variable será muy distinto.

Una primera clasificación simple es la expuesta anteriormente entre las variables cuyos valores son cualidades o categorías, también llamadas atributos, y aquellas cuyos valores son números con propiedades aritméticas. La edad y el sexo son ejemplos claros de ambos tipos de variable. Pero también lo son la clase social (con sus distintas categorías) y los ingresos (expresados en dólares, pesetas o euros; pero, en todo caso, cantidades).

A su vez, entre las variables cualitativas se distinguen las *nominales*, cuyos valores sólo poseen la propiedad de la identidad (cualquier valor es igual a sí mismo y diferente del resto) y las *ordinales*, en las que puede establecerse una jerarquía completa entre valores de manera que

² Cabe recordar que los nombres de las variables en el programa SPSS han de tener como máximo ocho caracteres (letras o números) contiguos de los que el primero no puede ser número.

si un valor llamado a está situado antes de un segundo denominado b , a su vez, éste precede a un tercero, al que se conocerá con c , necesariamente el primero ha de estar ubicado por delante del tercero. Ambas propiedades pueden formularse como sigue:

Principio de identidad:

$$\begin{array}{l} a = a \\ a \neq b \end{array} \quad (1)$$

Propiedad ordinal de los valores:

$$(a < b) \wedge (b < c) \rightarrow (a < c) \quad (2)$$

A su vez, las variables cuantitativas pueden clasificarse en variables de intervalo o de razón según carezcan o tengan un valor 0 que represente la ausencia total de la calidad que están representando. El cociente intelectual sólo puede ser clasificado de variable de intervalo pues el valor 0 es arbitrario y no equivale a la carencia absoluta de inteligencia; en cambio, puede catalogarse como variable de razón a los ingresos medidos, por ejemplo, en euros ya que en este caso el cero indica la ausencia total de lo que expresa la variable. No se trata, como a veces suele confundirse, de que la variable tenga o no el valor cero para catalogarla de una u otra forma, sino del significado que tiene este valor.

Otra clasificación útil para variables cuantitativas es la que separa a las variables discretas de las variables continuas. Teóricamente, las primeras son aquellas con limitado número de valores de modo que entre dos valores contiguos es imposible encontrar empíricamente un tercero con un valor intermedio. Una persona puede tener dos o tres hermanos; pero no dos hermanos y medio. En cambio, en las variables continuas siempre será posible imaginar valores intermedios pues el número de ellos es infinito. Así, entre una persona que pesa 60 Kg y otra que pesa 61 Kg, es posible encontrar otra con 60,5 Kg; la única limitación estaría en la precisión de los instrumentos de medida.

Otra clasificación del tipo de variable es la que se utiliza para su introducción en una matriz de datos informática. Desde este punto de vista, en los programas estadísticos se pueden distinguir entre variables *numéricas*, variables *cadena* y variables *fecha*. Estas últimas constan de tres espacios para la introducción respectiva del día, mes y año, las variables cadena son de tal naturaleza que pueden introducirse en ellas tanto números como caracteres. En cambio, las variables numéricas solo aceptan caracteres numéricos. Aunque parece que haya una correspondencia entre variables cadena y numéricas, por un lado, y variables cualitativas y cuantitativas, por el otro, la equivalencia es meramente aparente, pues generalmente en la matriz informática se introducen los valores de las variables en formato numérico y, rara vez, se emplean

caracteres para representarlos³. De este modo, cada vez que hay que introducir el dato “mujer”, en lugar de grabarse literalmente estas cinco letras, se introduce un código numérico, el 2 por ejemplo, que representará al valor real denominado etiqueta, que en este caso es ser mujer.

La tabla de distribución de frecuencias

El tratamiento más simple que puede darse a una matriz de datos no difiere según se trate con una u otra variable. La forma más simple de resumir la información de un conjunto de datos es la tabla de distribución de frecuencias, que consiste en presentar para cada valor de una —y sólo una— variable el número (*frecuencia*) de casos que lo comparte. Siguiendo el ejemplo de la Tabla 1, de los cuatro casos presentes en la matriz de datos, tres son varones y uno una mujer. De igual modo, en la variable edad, los cuatro casos poseen valores distintos, cada uno de ellos, por tanto, con una frecuencia de una unidad. La disposición típica de una tabla de distribución de frecuencias consiste en:

- a) Exponer como encabezamiento el nombre de la variable.
- b) Listar en la primera columna el repertorio de los distintos valores que presenta la variable entre los sujetos en estudio.
- c) Mostrar en la segunda columna la frecuencia correspondiente a cada valor. Esta segunda columna se finaliza con el sumatorio de todas las frecuencias, lo que equivale a expresar el número total de casos analizados.
- d) Por último, es conveniente crear una tercer columna con las *proporciones o frecuencias relativas*, que consisten en el cociente entre las frecuencias simples y su sumatorio. Más útil aún es transformarlas en porcentajes, pues de esta forma son de más fácil interpretación y la comunicación con el lector o auditor resulta favorecida (tabla 2).

³ La excepción no siempre seguida a esta regla afecta a los identificadores de caso, como puede ser el nombre del país en el supuesto de tener una base de datos con sus indicadores socio-económicos. La razón de ello es que si cada caso tiene un valor distinto, el uso de etiquetas no es más simple que el de la introducción directa de valores. Aun así, también en estos casos, puede utilizarse una variable numérica con tantos valores como casos tenga la matriz de datos.

Tabla 2.- Distribución de frecuencias de la variable *sexo*.

SEXO			
		Frecuencia	Porcentaje
Válidos	Hombre	3	75.0
	Mujer	1	25.0
	Total	4	100.0

e) Los pasos c) y d) pueden abreviarse en uno solo, si se exponen junto con cada valor (b) sólo los porcentajes y en la última fila el número de casos entre paréntesis para indicar que ya no se trata de una medida porcentual. (Ver la tabla 3).

Tabla 3.- Distribuciones de frecuencias de las variables *sexo* y *edad*.

Variable: Sexo		Variable: Edad	
Valores	Porcentajes	Valores	Porcentajes
Varón	75.0%	5	25.0%
Mujer	25.0%	6	25.0%
	(4)	34	25.0%
		35	25.0%
			(4)

La tabla 3 contiene dos variables. De los cuatro sujetos en estudio, el 75% son varones y el 25% mujeres. En relación con la edad, cada persona tiene un valor distinto en la variable.

Poco frecuentemente se realiza un estudio estadístico con tan sólo cuatro casos. A veces, con poca precisión, la Estadística ha sido definida como la ciencia de los grandes números, porque generalmente trata de describir grandes conjuntos, aunque para ello no necesite disponer de los datos de todos y cada uno de sus elementos. Se denomina *población* a ese gran conjunto del que se desea obtener una información, mientras que recibe el nombre de *muestra* un subconjunto de esa población extraído con unas determinadas condiciones que aseguren que el análisis que se efectúe con sus datos no difiera excesivamente del que se hubiese realizado teniendo la información de toda la población.

El tamaño que han de tener las muestras depende principalmente de cuán homogénea u heterogénea sea la población y, en menor medida, del tamaño de ésta última. Sin embargo, empíricamente, podría decirse que muestras inferiores a los treinta casos son muy pequeñas, entre esa cantidad y los doscientos siguen siendo pequeñas, entre esta última cifra y los 800 son muestras medianas, normales las comprendidas entre los 800 y los 3000 casos y por encima de varios miles pueden calificarse las muestras de grandes.

La matriz de datos con las que obtener las tablas de distribución de frecuencias tiene por tanto tantas filas como casos tenga la muestra y tantas columnas como variables tenga la investigación en estudio. Tampoco es usual organizar una investigación con sólo dos variables, a menos que sean muy difíciles de medir. Por regla general un estudio comprende un mínimo de diez variables y un máximo, en ocasiones escasas, de varios miles.

En este tipo de tablas la notación que se emplea para designar a los valores es x_i , con f_i se denominan las frecuencias absolutas, las frecuencias relativas se reconocen por p_i y el número de casos se expresa bien con n si los datos corresponden a una muestra, o N si se trabaja con los datos de una población. Por último, I denota el número de valores distintos que posee la variable. Cuando los valores de una tabla son exhaustivos y mutuamente excluyentes son evidentes las siguientes igualdades:

$$\sum_i^I f_i = n$$

$$\sum_i^I p_i = 1$$
(3)

Otros porcentajes que pueden ser incluidos en una tabla son los porcentajes (sobre casos) válidos y, en el caso de que no se trate de variables nominales, los porcentajes acumulados. Los primeros son aquellos que se calculan con el número de casos de los que disponemos de información, en lugar de con el tamaño total de la muestra. Si de las cuatro personas de la familia, no se dispusiera del dato de la edad de la madre; entonces podría calcularse dos porcentajes: uno sobre las cuatro personas que se sabe componen la familia y otro sobre las tres personas de las que se dispone información. En el primero, el 25% de los miembros de la familia tienen 35 años (la edad del padre); en el segundo el 33% de los sujetos de los que tenemos información tienen 35 años. Para formular este porcentaje, se considerará I' como el número de valores distintos y no perdidos que tiene la variable y como n' el número de sujetos con valores conocidos. De esta forma,

$$n' = \sum_{i=1}^{I'} f_i$$
(4)

y los porcentajes (p_i') de los I' valores se calculan dividiendo por el nuevo tamaño muestral (n').

$$p_i' = \frac{f_i}{n'} \quad (5)$$

Los otros porcentajes que pueden calcularse son los *acumulados* (P_i), que se obtienen sumando progresivamente los porcentajes de los valores menores (o mayores) de los que se desean obtener. Generalmente, se acumulan sobre los porcentajes válidos, en lugar de los porcentajes simples.

$$P_i = \sum_{i=1}^i p_i \quad (6)$$

De este modo, la tabla edad quedaría completa con sus frecuencias y tres tipos de porcentajes:

Tabla 4.- Distribución de frecuencias de la variable *edad*.

EDAD					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	5	1	25.0	25.0	25.0
	6	1	25.0	25.0	50.0
	34	1	25.0	25.0	75.0
	35	1	25.0	25.0	100.0
	Total	4	100.0	100.0	

La tabla 4 muestra una simple distribución de cuatro casos en la variable edad. Cada valor tiene una frecuencia única, por lo que existe un 25% de casos con 5 años, otro 25% con 6, 25% con 34 y, finalmente, otro cuarto que ha cumplido 35 años. Los porcentajes acumulados centrales, los únicos que merecen comentarse, señalan un 50% de personas con 6 o menos años y un 75% con la edad de 34 años o inferior.

Estadísticos resúmenes de distribuciones

Las distribuciones son un resumen de los datos disponibles de las muestras generalmente, pues pocas veces se cuenta con los datos de la población. Aun con todo, es posible condensar aún más la información con la ayuda de los estadísticos, que son datos calculables en la distribución que informan de alguna de sus características. Cinco son las principales

características que pueden resumirse en una distribución: la tendencia central, la posición, la dispersión, la simetría y el apuntamiento.

Medidas de tendencia central

Por tendencia central se entiende un valor que representa al conjunto de valores de la distribución de una variable. En el caso extremo de una distribución en la que todos los sujetos tuvieran el mismo valor, ese dato daría cuenta de todos ellos. Pero, como su propio nombre indica, las variables no se caracterizan por presentar valores únicos. Por ello, hay diversos procedimientos para obtener una medida de tendencia central. Las más conocidas y empleadas son:

a) La moda: valor que posee la mayor frecuencia de una distribución. Si en un grupo de cinco personas, tres son varones y dos mujeres; la moda es ser hombre. En la distribución de la tabla 4, donde hay cuatro casos, no existe moda porque los cuatro valores poseen la misma frecuencia. Para que haya moda, ha de existir un valor con mayor frecuencia que el resto⁴.

b) La mediana es el valor que ocupa la posición central de una distribución ordenada por sus valores. En consecuencia, no tiene sentido su cálculo en el caso de variables nominales. Para obtenerla hay que buscar en una tabla de distribución de frecuencias el primer valor cuya frecuencia acumulada supere el 50%. Así, si se dispone de tres valores {4, 7, 6}, la mediana es 6, pues previamente ordenados, es el que ocupa el medio de la distribución y es el primero cuya frecuencia acumulada (66.6%) está por encima del 50%. En la tabla 4, la mediana corresponde a dos valores, pues posee un número par de casos. Por convención, se adopta que la mediana sea la semisuma de los dos valores centrales. En este caso, $(6+34)/2$, es decir, 20. Por tanto, para obtener la mediana cuando un determinado valor posea una frecuencia acumulada igual al 50%, es preciso calcular la semisuma con el siguiente valor de la tabla.

También puede calcularse la mediana obteniendo la posición central u ordinal de la mediana $O(Me)$ y buscando el valor que ocupa la posición hallada mediante la siguiente fórmula:

$$O(Med) = \frac{n+1}{2} \quad (7)$$

c) La tercera medida de tendencia central es la media aritmética, que es un promedio de los valores de la distribución obtenido mediante la división de la suma de todos los valores por el número de casos. La cantidad ofrecida por la media es, utilizando un aforismo, el valor que tendrían todos los valores en el supuesto de que todos los valores tuvieran el mismo valor. Si en un grupo humano una persona tiene un hermano, otra dos y la tercera tres, poseen en total seis

⁴ No obstante, cuando hay dos valores con la misma frecuencia, se puede hablar de distribuciones bimodales, lo que implica la existencia de dos modas.

hermanos, que si se distribuyeran equitativamente corresponderían a dos por persona. La obtención de este estadístico responde a la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^I x_i f_i}{n} \quad (8)$$

Así la media de edad en la familia del ejemplo considerado sería de 20 años, que es el cociente entre la suma de las edades (80) de las cuatro personas y el número de miembros que la componen (4).

Medidas de localización

Son medidas de localización aquellas que indican el valor que ocupa un determinado orden en una distribución. La medida más simple de localización es la también medida de tendencia central, la mediana, pues es el valor que ocupa la posición del centro de la distribución, o dicho de otro modo, el 50% de las observaciones de la distribución tiene valores menores o iguales a de ella y el otro 50% tiene valores mayores o iguales. La mediana también puede ser concebida como aquel valor que divide a la distribución en dos partes iguales.

Otras medidas de localización son los cuartiles, que pueden ser definidos como tres valores que dividen a la distribución en cuatro partes iguales. Así, el primer cuartil tiene un 25% de casos por debajo de dicho valor; el segundo cuartil coincide con la mediana y el tercero presenta un 25% de casos con valores superiores. Para obtenerlos, se calcula, en primer lugar, las posiciones de los cuartiles - $O(Q_1)$ y $O(Q_3)$ - y a partir de ellas se extraen los valores correspondientes. Las posiciones respectivas del primer y tercer cuartil (el segundo es igual a la mediana) son:

$$\begin{aligned} O(Q_1) &= \frac{n+1}{4} \\ O(Q_3) &= \frac{3(n+1)}{4} \end{aligned} \quad (9)$$

Una vez obtenidas las posiciones, se buscan los valores que las ocupan. En el caso de que $O(Q_x)$ dé un valor decimal, se procede a una suma de los valores que ocupan la parte entera de la posición y el que ocupa la siguiente multiplicados respectivamente por (1-fracción) el primero y por la fracción decimal el segundo.

Así, en el ejemplo de la tabla 4, dado que son cuatro casos, al primer cuartil le correspondería la posición 1.25 y al tercero la 3.75. En consecuencia, el valor del primer y tercer cuartil serían respectivamente de

$$Q_1 = 5 \times (1 - 0.25) + 6 \times 0.25 = 5.25$$

$$Q_3 = 34 \times (1 - 0.75) + 35 \times 0.75 = 34.75$$

De similar familia son los deciles y percentiles. En el primer caso, son nueve valores que dividen a la distribución en diez partes iguales y, en el segundo, noventa y nueve que parte los datos en cien subconjuntos del mismo tamaño.

Para hallar lo n-til, se procede de modo similar a cuando se obtienen los cuartiles. Se busca la posición correspondiente al n-til y si ésta es decimal, se interpola los dos valores contiguos con los pesos de la fracción complementaria y original⁵. En general, la posición de un n-til (T_x) se ajusta a la siguiente fórmula:

$$O(T_x) = \frac{x(n+1)}{T} \quad (10)$$

De este modo, el quinto sextil de una distribución con 35 casos, ocuparía la posición trigésima: $5(35+1)/6$.

En el análisis exploratorio, en lugar de utilizar cuartiles, deciles y percentiles, se emplean medidas similares -aunque no siempre iguales a ellos- que son los cuartos (F_i y F_s)⁶ o bisagras (*hinges*) propuestos por Tukey (1977), los octavos (E_i y E_s), los dieciseisavos (D_i y D_s), los treintadosavos (C_i , C_s), ... Definidos operacionalmente como la mediana de las medidas de localización precedentes, lo que les diferencia es la posición en la que se buscan, pues para los cuartos viene definida por las siguientes expresiones:

$$O(F_i) = \frac{[O(Med) + 1]}{2} \quad (11)$$

$$O(F_s) = n + 1 - O(F_i)$$

En el supuesto de que n sea igual a 4, por ejemplo, la $O(Me)$ es 2.5, los valores de $O(F_i)$ y $O(F_s)$ son respectivamente 1.5 y 3.5 y, aplicando la misma regla de la diferencia de fracciones mostrada en (10), se obtiene los valores de los cuartos:

$$F_i = 5 \times (1 - 0.5) + 6 \times 0.5 = 5.5$$

$$F_s = 34 \times (1 - 0.5) + 35 \times 0.5 = 34.5 \quad (12)$$

5 Cuando los datos de la distribución están agrupados en intervalos, existe también una fórmula para estimar el valor exacto dentro del intervalo, pero su explicación queda fuera del alcance de este texto.

6 Reciben la denominación de F de la palabra inglesa *forth*, los subíndices i y s indican respectivamente los cuartos inferior y superior. Se respeta la letra F para continuar la serie de estadísticos de localización exploratorios E (*eighth*), D, C, B, .. de las distribuciones.

Como puede apreciarse son ligeramente distintos de los cuartiles.

La posición de los octavos, a su vez, queda definida como:

$$\begin{aligned} O(E_i) &= \frac{[O(F_i) + 1]}{2} \\ O(E_i) &= n + 1 - O(E_i) \end{aligned} \quad (13)$$

En la distribución de cuatro casos del ejemplo seguido, como $O(F_i)$ es igual a 1.5, las posiciones de los octavos inferior y superior son las de 1 y 4, es decir, las correspondientes a la primera y última observación, cuyos valores respectivos son 4 y 35 años.

Para ofrecer una visión global de la distribución, estas medidas se disponen en una tabla de posición del siguiente modo:

Tabla 5.- Tabla de posiciones de la variable edad (n=4).

Me	20	
F	55	345
E	5	34

En esta tabla se advierte cómo en el centro de la distribución hay una diferencia de edades considerable, mientras que en los extremos los años de las personas son muy similares. En este caso, la función de esta tabla es relativa, en la medida que hay en ella más valores que casos, con lo que resultaría más pertinente obtener una idea precisa de la distribución mediante el examen de la distribución original.

En cambio, para examinar la distribución de la edad en una muestra de 1200 casos con unos 70 valores distintos, por ejemplo, la tabla de posiciones da una idea más certera de la distribución que la mera tabla de frecuencias. En la próxima tabla se muestra cuán pocos casos existe en el extremo superior de la distribución, en contraste con los casos con el valor mínimo (18 años). Además, el que los valores de la izquierda estén más próximos a la mediana, indica una mayor proporción de casos en cada uno de ellos que en cada uno de los valores superiores a la mediana.

Las tablas de posiciones pueden ampliarse con otro estadístico: la semisuma de las medidas de posición (desde el cuarto hasta el n-avo), que en realidad son medidas de posición central alternativas que se denominan los centros de los cuartos, octavos, dieciseisavos, ...

Tabla 6.- Tabla de posiciones de la variable edad (n=1200).

Me	44		
F	29	455	62
E	23	455	68
D	20	475	75
C	18	485	79
B	18	50	82
A	18	51	84
Z	18	515	85
Y	18	535	89
X	18	545	91
1	18	55	92

Asimismo pueden considerarse medidas de localización los valores mínimo y máximo, pues son aquellos que se ubican en la primera y última posición de la tabla; pero también, en el análisis exploratorio se habla de los casos atípicos, que son los que están alejados de un cuarto más de vez y media la distancia entre cuartos. De la tabla 6 se deduce que no hay valores atípicos por debajo de la distribución pues, siendo la distancia entre cuartos de 33 años, no hay caso con una edad inferior a 29 (cuarto inferior) menos $1,5 \times (62-29)$ años. Ni tampoco los hay por encima del cuarto superior porque para ello tendrían que tener más de 102 años ($62+39,5$). Y, de la misma forma, se habla de casos extremos cuando están alejados más de tres veces la distancia entre cuartos. Es evidente que tanto los valores atípicos como los extremos de una distribución no son únicos, sino más bien rango de valores.

En una distribución con los siete valores siguientes: {10, 24, 26, 30, 30, 32, 60}, el valor 10 sería atípico, pues el cuarto inferior es 25 y el superior 31; por tanto la distancia entre ellos es de 6, y el valor de 10 se encuentra a una distancia superior a 9 ($1,5 \times 6$) del cuarto inferior. Por otro lado, el 60 es un caso extremo pues el cuarto superior (31) mas 3 veces la distancia entre cuartos (3×6) no llega a superarlo.

Medidas de dispersión

El tercer tipo de medidas son las llamadas medidas de dispersión. Indican cuán alejados están los valores de la distribución de aquel valor que los representa. Los estadísticos de dispersión más utilizados son:

a) La dispersión modal: es la proporción (o porcentaje) de sujetos de una distribución que no tienen el valor modal. Este simple estadístico es uno de los escasos que se pueden utilizar para

obtener la dispersión en variables nominales u ordinales. Su fórmula se representa del siguiente modo:

$$D_{mo} = 1 - p_{mo} \quad (14)$$

Así basta restar a uno la proporción de casos que tienen la moda. En el ya conocido ejemplo del grupo de cuatro personas, tres de las cuales son hombres, la dispersión modal sería del 25%, pues esta es la proporción de personas que no son varones.

b) El rango: es la diferencia entre los valores extremos de una variable. En el caso de la variable edad en la familia de cuatro miembros que sirve de ejemplo, el rango toma el valor de 30 años, pues es la diferencia entre la edad (35) del mayor y la del menor (5). Esta medida puede estar muy condicionada por un solo valor extremo poco representativo de lo que se estudia. Imagínese un grupo de 200 personas de edades comprendidas entre 17 y 18 años, salvo una que tiene 60. En este caso decir que el rango es de 43 años daría una imagen sesgada de este agregado. Por ello, se utiliza frecuentemente el llamado rango intercuartílico, que es la diferencia entre los valores correspondientes al tercer y primer cuartil. Así en el caso de la familia, sería de 28 años, mientras que en los dos centenares de personas el rango intercuartílico sería de 1 año.

c) Del mismo modo que existe un rango intercuartílico, el análisis exploratorio emplea la amplitud intercuartiles, interoctavos, interdieciseisavos, y así sucesivamente. Fruto de ello, se puede ampliar la tabla de posiciones, con los rangos de cada una de estas posiciones. A la tabla 6 habría que añadirle una columna con las correspondientes amplitudes.

Lo que muestran las sucesivas amplitudes es cómo se va pasando desde la amplitud que abarca la mitad de los datos (33), hasta el rango que abarca al conjunto de ellos (74).

Tabla 7.- Tabla de posiciones de la variable edad (n=1200).

Me	44			Amplitud
F	29	455	62	33
E	23	455	68	45
D	20	475	75	55
C	18	485	79	61
B	18	50	82	64
A	18	51	84	66
Z	18	515	85	67
Y	18	535	89	71
X	18	545	91	73
1	18	55	92	74

d) La desviación media es un promedio de los valores absolutos de las desviaciones de los valores con respecto a la media aritmética. Ha de advertirse que se trata de promedio de valores absolutos, pues si no se prescindiese del signo de las desviaciones, por una importante propiedad de la media aritmética, siempre arrojaría el valor de cero. En la distribución de la edad de los miembros de la familia hay cuatro desviaciones sobre la media (20 años): el hijo menor se desvía -15 años de la media, el hijo mayor -14 años, la madre 14 años y el padre 15 años. La suma de estas cuatro desviaciones es 0, a menos que se añadan los valores sin considerar el signo que les precede, en cuyo caso la suma es de 58 años. De ahí se obtiene el promedio con la división de esta cantidad entre las cuatro personas que componen las observaciones realizadas, 14'5, que representa lo que se desvía en promedio cada caso de la media aritmética. Ello es obvio pues dos casos se alejan de la media en 14 años y otros dos en 15 años.

e) Otro estadístico de dispersión importante para el análisis exploratorio de datos es la mediana de las desviaciones absolutas (MeDA). Tiene la ventaja de que, siendo un promedio que abarca todas las puntuaciones, no se ve afectada por los valores extremos como otras medidas que utilizan la media aritmética. Su fórmula es:

$$MeDA = Med |x_i - Med| \quad (15)$$

En el ejemplo seguido con cuatro casos {5, 6, 34, 35} y mediana igual a 20, el valor que adopta es el 14.5, pues las desviaciones absolutas son {15, 14, 14, 15} y la mediana de ellas es la cantidad mencionada.

f) La varianza es una media aritmética de las desviaciones cuadradas de los valores con respecto a la media. En lugar de promediar los valores absolutos de las desviaciones, éstas se

elevan al cuadrado para que su suma no sea cero y, de este modo, se penalizan las desviaciones más alejadas de la media. Así, el cuadrado de una unidad de desviación sigue siendo uno, el de dos desviaciones es cuatro, el de tres, nueve; el de 10, 100 y así sucesivamente va aumentando en progresión geométrica a medida que las desviaciones se hacen mayores.

En la distribución de la edad en la familia de ejemplo, tanto el padre como el hijo menor se desvían 225 años cuadrados de la media y la madre y el hijo mayor 196 cada uno de ellos. En consecuencia, el promedio de años cuadrados que se desvían estos cuatro sujetos de la media de 20 es de 210'5. Esta operación se formula del siguiente modo:

$$s^2 = \frac{\sum_{i=1}^I (x_i - \bar{x})^2 f_i}{n} \quad (16)$$

e) La desviación típica es la raíz cuadrada de la varianza. Se utiliza para devolver el valor de la varianza a sus unidades originales. Como acaba de verse, la varianza de 210'5 está referida en años cuadrados. Para poder hablar en términos de años, hay que hallar la raíz cuadrada de este valor, resultando ser de 14,51 años⁷. Su cálculo se obtiene mediante la expresión:

$$s = \sqrt{\frac{\sum_{i=1}^I (x_i - \bar{x})^2 f_i}{n}} \quad (17)$$

f) El coeficiente de variación es una medida de dispersión relativa. Es el cociente entre la desviación típica y el valor absoluto de su correspondiente media aritmética. Al ser una razón o cociente, carece de unidades y, en consecuencia, se utiliza para comparar la dispersión entre variables que tengan distintas unidades de medida. Como la varianza y la desviación típica son siempre positivas, este coeficiente tampoco tiene sentido que sea negativo, aunque la media posee valores negativos. Por tanto su valor es 0, como el de las dos medidas precedentes, en el caso de que todos los valores de la variable sean idénticos y, salvo distribuciones muy dispersas, es inferior a 1.

$$CV = \frac{s}{|x|} \quad (18)$$

7 Nótese que la desviación típica es siempre mayor o igual que la desviación media habiendo tanta mayor diferencia entre ellas cuanto mayor sea la dispersión en la variable.

g) Desde el punto de vista exploratorio, se utiliza otra medida de dispersión relativa basada en los cuartos. Se trata del coeficiente de variación intercuartílico, que se obtiene mediante el cociente de la diferencia de cuartiles y su suma:

$$CVI = \frac{Q_3 - Q_1}{Q_3 + Q_1} \quad (19)$$

Este coeficiente tiene la ventaja de su mayor resistencia a la influencia de casos extremos. Es de gran utilidad para la comparación de la dispersión entre distintas distribuciones. A semejanza del coeficiente de variación, este caso es el cociente entre una dispersión (el rango intercuartílico) y un promedio, el de cuartiles. Una modalidad algo distinta, pero con resultados muy similares, es aquella que usa los cuartos en lugar de los cuartiles.

Medidas de simetría

Existen otras medidas cuyo propósito es expresar a través de un número la forma de la distribución. Estas se clasifican, a su vez, en dos tipos, las de simetría (que atienden a la forma horizontal de la distribución: si la izquierda de la distribución es semejante a su derecha) y las de apuntamiento (que expresan a la distribución vertical de los valores: si las frecuencias de los valores centrales son mayores que las de los valores extremos).

Para variables continuas, existe un patrón o modelo de distribución de la estadística llamado distribución normal que se caracteriza por: a) tener idéntica la media, la moda y la mediana, b) ser simétrica, es decir, la distribución de los valores por debajo de la media es refleja de la distribución de los valores por encima de la media idéntica y c) poseer un alto número de casos en los valores centrales e ir descendiendo esta frecuencia a medida que los valores se van alejando del centro de la distribución, esto es, de la mediana.

Las dos primeras propiedades están muy ligadas entre sí, pues en toda distribución simétrica unimodal, tienen los mismos valores los tres estadísticos de tendencia central. Además, se sabe que la relación entre ellos tiende a ser empíricamente la siguiente:

$$3(\bar{x} - Med) \approx \bar{x} - Mo \quad (20)$$

De este modo, se sigue que utilizando cualquiera de los dos términos de la ecuación (20), se obtiene una medida de la simetría de la distribución de tal suerte que un valor de 0 indica la presencia de una distribución simétrica, un valor positivo indica una asimetría en la que la media se encuentra por encima (a la derecha) de la posición central y un valor negativo evidencia una posición de la media por debajo (a la izquierda) de la moda y la mediana.

Véase ello con ejemplos de tres distribuciones muy simples de cinco casos, en los que los tres centrales son igual a cinco:

a) {4, 5, 5, 5, 6}. Med=5; Mo=5; \bar{x} =5; (\bar{x} -Mo=0).

b) {3, 5, 5, 5, 6}. Med=5; Mo=5; \bar{x} =4.8; (\bar{x} - Mo =-0.2).

c) {4, 5, 5, 5, 7}. Med=5; Mo=5; $\bar{x}=5.2$; (\bar{x} - Mo = 0.2).

La distribución a) es simétrica: por debajo de los cinco hay un valor que se desvía un punto negativo, mientras que por encima de ellos hay también un único valor que se desvía un punto positivo. La distribución b), sin embargo, pesa más a la izquierda, pues el valor 3 se aleja un punto más de la media que el valor 6. Y en la distribución c) ocurre lo contrario, el valor 7 está más alejado del centro que el valor 4.

Sin embargo, para poder comparar la simetría entre distribuciones con distintas unidades de medida, se utiliza la desviación típica. De este modo la fórmula quedaría en cualquiera de las dos siguientes modalidades:

$$As = \frac{\bar{x} - Mo}{s} = \frac{3(\bar{x} - Med)}{s} \quad (21)$$

También es posible obtener un índice de simetría mediante el momento de orden 3 con respecto a la media, esto es, el promedio de las desviaciones cúbicas de los valores y su media:

$$m_{\bar{x}}^3 = \frac{\sum_{i=1}^I (x_i - \bar{x})^3 f_i}{n} \quad (22)$$

La fórmula del momento es de tal naturaleza que si hay predominio de valores por debajo (a la izquierda) de la media ($x_i - \bar{x}$), sale negativo y si hay predominio de valores por encima, resulta positivo. También para obtener un coeficiente de simetría con el que poder hacer comparaciones entre variables se divide este momento de orden 3, cuyas unidades son cúbicas, por la desviación típica al cubo:

$$As = \frac{m_{\bar{x}}^3}{s^3} \quad (23)$$

Desde la óptica del análisis exploratorio, ambos coeficientes de asimetría son poco resistentes a los valores extremos, por ello se prefiere utilizar otros coeficientes que eviten la influencia de esos valores.

El que en mayor medida evita la influencia de los casos extremos es el índice de Yule pues sólo tiene en cuenta los valores de los cuartos. Se basa en el hecho de que en una distribución simétrica el centro de los cuartos es igual a la mediana, mientras que si es mayor es porque ambos cuartos se desplazan hacia la derecha de la distribución y si es menor es porque se han desplazado hacia la izquierda.

$$H_1 = \frac{F_1 + F_3 - 2Med}{2Med} \quad (24)$$

En la distribución de los cuatro casos de la familia el valor de H_1 es cero, pues es una distribución simétrica; en cambio, la distribución de la edad en la muestra de 1200 casos es asimétrica hacia la derecha pues hay mayor frecuencia en las edades bajas que en las edades altas ($H_1=0,03$). Sin embargo, como el cálculo se centra entre las edades de 29 y 62 años (valores de los cuartos), el valor del coeficiente arroja una distribución casi simétrica.

Por ello, Kelley propone otro en el que se trabaje con los deciles primero y noveno, restando a la mediana su promedio:

$$H_2 = Med - \frac{D_1 + D_9}{2} \quad (25)$$

Su principio es el mismo que el H_1 ; a medida que los valores del decil superior son mayores, la simetría se deriva hacia la derecha y a la inversa con los deciles inferiores. Sin embargo, arroja valor negativo para distribuciones asimétrica a la derecha y positivo para las escoradas a la izquierda. Para equipararse con H_1 , conviene utilizar otro coeficiente con relación lineal con H_2 , el H_3 :

$$H_3 = \frac{D_1 + D_9 - 2Med}{2Med} = -\frac{H_2}{M} \quad (26)$$

El valor que adopta en la distribución de 1200 casos es de 0.06 pues los deciles primero y novena se encuentran en los valores 22 y 71.

Medidas de apuntamiento

La otra medida sobre la forma de la distribución es el apuntamiento, que indica cuán centradas o dispersas están las frecuencias de los valores en relación con el punto medio de la distribución. Si las frecuencias están concentradas en el centro, entonces la distribución se llamará leptocúrtica, si las frecuencias mayores se ubican en los extremos de la distribución, la distribución será platicúrtica y, en el caso intermedio, sería una distribución mesocúrtica. Para verlo más claramente, véanse estas cinco distribuciones:

- a) {1, 1, 5, 9, 9}. $\bar{x}=5$; $s=0.9$; $K=-3$.
- b) {4, 4, 5, 6, 6}. $\bar{x}=5$; $s=3.6$; $K=-3$.
- c) {2, 4, 5, 6, 8}. $\bar{x}=5$; $s=2.0$; $K=0.2$.
- d) {2, 5, 5, 5, 8}. $\bar{x}=5$; $s=1.9$; $K=2.0$.

Así se ve que la distribución a y b, aun con desviaciones típicas distintas, poseen la misma curtosis pues en ambas hay mayor frecuencia en los valores extremos que en el central. La curtosis negativa indica una distribución platicúrtica. La distribución c) en cambio es casi mesocúrtica ($k=0.2$) porque tiene extendidas sus frecuencias de modo que hay un número medio en el centro (del 4 al 6) y un número inferior en los extremos; por último, la distribución d) es

leptocúrtica (positiva) porque la frecuencia de los valores de la distribución están concentrada en el centro (tres valores en la media, frente a uno en los extremos).

Para calcular la curtosis se utiliza el momento de orden 4 con respecto a la media dividido, para que quede desprovisto de unidades, por la desviación típica a la cuarta. Además a este cociente se le resta tres unidades para que este estadístico arroje un valor de cero en el caso de que se trate de una distribución normal.

$$k = \frac{m_x^4}{s^4} - 3 = \frac{\sum_{i=1}^I (x_i - \bar{x})^4}{n s^4} - 3 \quad (27)$$

También es preferible desde la óptica del análisis exploratorio utilizar otro índice para expresar el apuntamiento de una distribución. Para ello se compara la dispersión existente en el 90% de los casos centrales, con la existente en el 75%. Un mayor rango interdecílico, en relación con la amplitud intercuartiles, indica una menor presencia de casos extremos, por lo que la distribución con un relativamente amplio recorrido interdecílico serán leptocúrticas. En el caso de las distribuciones normales la razón entre ambos rangos es de 1.96. En consecuencia en la expresión:

$$K_2 = \frac{D_9 - D_1}{1.9(F_4 - F_3)} \quad (28)$$

proporciona valor 1 en el caso de distribuciones mesocúrticas, un valor superior a la unidad en las distribuciones leptocúrticas e inferior en el de las platicúrticas. En el caso de la edad en la muestra representativa de la población española, el valor de K_2 es igual a 0.78.

Los estimadores robustos centrales

En la introducción de este libro se dijo que la resistencia era una de las cualidades más importante en el trabajo exploratorio con los datos. Se definió como la insensibilidad de un estadístico ante el cambio de un parte pequeña del conjunto de datos.

En las medidas de tendencia central se vieron tres: la moda, la mediana y la media aritmética. La moda no es un estimador resistente, pues un solo caso podría cambiarla desde el valor mínimo hasta máximo. En una muestra con tres casos $\{1, 1, 5\}$, basta con cambiar uno de los unos por un cinco para que la moda pase desde el menor valor hasta el mayor de la distribución. En el caso de la media, un cambio en el valor de un caso también altera sustancialmente el promedio. En la distribución anterior, si el 5 fuera 2997, el promedio, pasaría a ser de 1000. La mediana, en cambio, es más resistente a los cambios en subconjuntos pequeños de datos. Aunque con tres

casos sea similar a la moda, a medida que se acrecienta el n , las ventajas de la mediana en resistencia son claras pues siempre que no haya un cambio de datos de un lado de la distribución al otro, la mediana no queda modificada.

La media recortada, estimador más resistente que la media aritmética, consiste en el promedio de los $(1-2\alpha)$ valores centrales de una distribución. Así, en una distribución con ocho casos la media recortada del 25%, implica el promedio de los cuatro casos centrales, descartando los valores mínimo y máximo de la variable.

Una variable recortada puede ser considerada como una media ponderada. En general, se llama media ponderada porque en su cálculo se da distinto peso (w_i) a los valores de la distribución. En cualquier situación, se obtiene la media ponderada mediante la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i} \quad (29)$$

La media recortada, considerando que se cuentan los valores centrales (teniendo como peso la unidad) y se descartan los extremos (peso nulo),

$$w_i = \begin{cases} 0 & \text{si } i \leq \alpha n \text{ o } i \geq n+1 - \alpha n \\ 1 & \text{si } i > \alpha n \text{ y } i < n+1 - \alpha n \end{cases} \quad (30)$$

puede ser vista como un caso especial de media ponderada, en la que los casos centrales tienen una ponderación igual a la unidad y los casos extremos una ponderación igual a 0.

Por robustez se entiende en estadística la calidad de un estimador en tanto y cuanto no se altera su eficiencia por cambios en las características de los datos. Lo que caracteriza a un estimador robusto es su insensibilidad a desviaciones de los datos de los modelos probabilísticos.

Los estimadores M son una familia de medidas de tendencia central que se caracteriza por minimizar las funciones de desviación de las observaciones de la estimación. La media y la mediana pertenecen a esta clase de estimadores pues cumplen con las siguientes condiciones de minimización:

La función objetiva de la media aritmética es

$$\rho(x_i, \bar{x}) = (x_i - \bar{x})^2 \quad (31)$$

Con ello, se quiere decir que el promedio \bar{O} es aquel que minimiza las desviaciones cuadradas de los valores en relación con el propio promedio. La fórmula de la media se obtiene precisamente diferenciando esta función objetiva y buscando el valor que soluciona la igualdad de la derivada y cero.

La mediana, por su parte, también es un estimador de la clase M; pero en su caso la función objetiva es

$$\rho(x_i, Med) = |x_i - Med| \quad (32)$$

Mediante otras funciones objetivas, que dan lugar a ponderaciones de las puntuaciones de los objetos distintas según lo alejado o no que se encuentren de un determinado promedio, se hallan estimadores más robustos que la mediana.

Además de la media y la mediana, los estimadores M más conocidos son los de Huber, Tukey, Andrews y Hampel.

Los pesos respectivos que hay que otorgar a los valores en cada uno de estos estimadores son los siguientes:

Estimador	Condición	w_i
Huber (k)	$ \mu \leq k$	1
	$ \mu > k$	$k \operatorname{sgn}(\mu)/\mu$
Hampel (a, b, c)	$ \mu \leq a$	1
	$a < \mu \leq b$	$a \operatorname{sgn}(\mu)/\mu$
	$b < \mu \leq c$	$a \frac{c - \mu }{c - b} \frac{\operatorname{sgn}(\mu)}{\mu}$
	$ \mu > c$	0
Andrews ©	$ \mu \leq 1$	$\operatorname{sen}(B\mu)/B\mu$
	$ \mu > 1$	0
Tukey ©	$ \mu \leq 1$	$(1 - \mu^2)^2$
	$ \mu > 1$	0

El cálculo de estos estimadores se realiza mediante un procedimiento iterativo. Se parte de un valor inicial de μ_i equivalente a la siguiente expresión:

$$\mu_i = \frac{x_i - x^*}{cMeDA} \quad (33)$$

donde x^* es una primera estimación del promedio, la media aritmética, por ejemplo.

A continuación se calcula un nuevo \bar{x}^* utilizando los pesos (w_i) que se han calculado con los u_i iniciales.

Con el nuevo \bar{x}^* , se vuelve a estimar las u_i según la fórmula (33) y con ellas se vuelve a calcular otra \bar{x}^* , hasta que converja con la anterior.

En la práctica, en el análisis exploratorio la mejor medida de centralidad es la mediana por cuanto su fácil cálculo y su alta resistencia proporciona suficientes garantías de exactitud. No obstante, si se precisa mayor acierto en la estimación y la distribución de la población es desconocida, debe elegirse cualquiera de los estimadores robustos acabados de explicar. Finalmente, la media aritmética, excepto en distribuciones con número considerable de casos extremos, debería usarse con especial cuidado y sólo cuando lo requieran los análisis posteriores como el análisis de varianza o la regresión lineal.

Obtención de las distribuciones y sus medidas.

Para la construcción de tablas de distribuciones de frecuencias hay dos posibles instrucciones: la más simple y directa con la instrucción *Estadísticas/Resumir/Frecuencias*. Para obtener la tabla basta con pasar a la derecha las variables cuya distribución se desee y dejar marcado el cuadro con el texto *Mostrar tablas de frecuencias*. Los estadísticos que se pueden obtener son los mostrados en el cuadro de diálogo ubicado en la derecha de la figura.

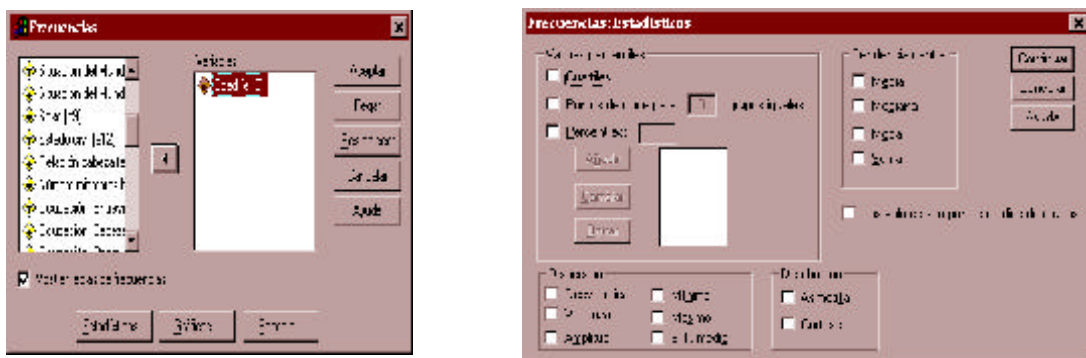


Figura 2.- Cuadro de diálogo principal y de estadísticos de la instrucción *Frecuencias*.

El otro método de obtención de frecuencias es mediante la instrucción *Estadísticas/Tablas personalizadas/Tablas de frecuencias*. Si se opta por esta modalidad, las posibilidades de dar un formato distinto al resultado se acrecientan. En el cuadro de diálogo principal de esta modalidad se ha de ubicar al menos una variable en la casilla *Frecuencia para* y pueden alterarse los *estadísticos*, el *diseño*, el *formato* y los *títulos*. Lo más sustancial en el cambio serían los estadísticos, donde se puede elegir si mostrar las frecuencias ponderadas, sin ponderar y los porcentajes, además de especificar el número de decimales deseados y la presencia de totales.



Figura 3.- Cuadro de diálogo principal y de estadísticos de la instrucción *Tablas*.

Hay otras dos instrucciones para la obtención de medidas de distribución: la una más en la línea de la estadística clásica (*Estadística/Resumir/Descriptivos*) con la que, además de permitir la tipificación de las variables (véase capítulo 3), se pueden obtener los estadísticos mostrados en el cuadro de diálogo:

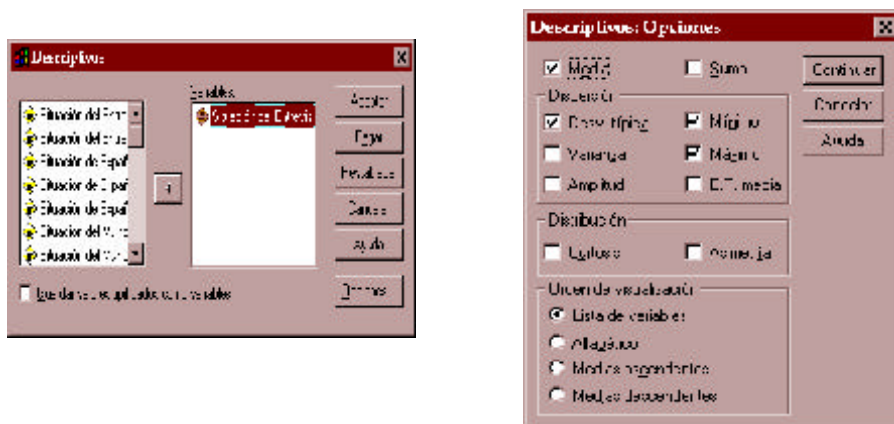


Figura 4.- Cuadro de diálogo principal y de estadísticos de la instrucción *Descriptivos*.

La otra instrucción está más vinculada con el análisis exploratorio (*Estadística/Resumir/Explorar*). Mediante el cuadro de diálogo de los estadísticos es posible solicitar los estadísticos descriptivos clásicos con su correspondiente intervalo de confianza, los estimadores M con sus correspondientes valores por omisión⁸ (Huber, 1.339; Tukey, 4.685; Hampel, 1.7, 3.4, 8.5 y

⁸ Estos pueden ser cambiados dando la orden mediante sintaxis en lugar de hacerlo por menú. Para más detalles consulte en el manual el comando *Examine*.

Andrews, 1.34B), los valores atípicos (los cinco⁹ con valores más extremos) y los percentiles (5, 10, 25, 50, 75, 90 y 95) acompañados por las bisagras de Tukey (Véase figura 5 y tabla 8).

Tabla 8.- Resultado de los programas *Tablas* y *Examinar* con

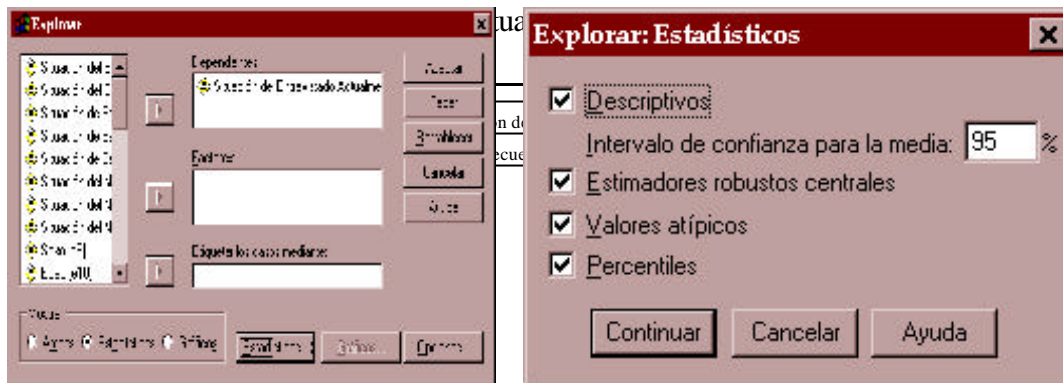


Figura 5.- Cuadro de diálogo principal y de estadísticos de la instrucción *Explorar*.

7	179	19.1%
8	104	11.1%
9	69	7.4%
10	63	6.7%
Total	936	100.0%

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Situación del Entrevistado en el Futuro	936	78.0%	264	22.0%	1200	100.0%

9 Este parámetro también puede ser cambiado introduciendo la instrucción mediante sintaxis.

Descriptivos

			Estadístico	Error típ.
Situación del Entrevistado en el Futuro	Media		6.34	6.66E-02
	Intervalo de confianza para la media al 95%	Límite inferior	6.20	
		Límite superior	6.47	
	Media recortada al 5%		6.41	
	Mediana		6.00	
	Varianza		4.155	
	Desv. típ.		2.04	
	Mínimo		0	
	Máximo		10	
	Rango		10	
	Amplitud intercuartil		3.00	
	Asimetría		-.443	.080
	Curtosis		.614	.160

Percentiles

	Percentiles	Valor	Bisagras de Tukey
Situación del Entrevistado en el Futuro	5	3.00	
	10	4.00	
	25	5.00	5.00
	50	6.00	6.00
	75	8.00	8.00
	90	9.00	
	95	10.00	

Estimadores-M

	Estimador-M de Huber ^a	Biponderado de Tukey ^b	Estimador-M de Hampel ^c	Onda de Andrews ^d
Situación del Entrevistado en el Futuro	6.30	6.36	6.41	6.36

a. La constante de ponderación es 1.339.

b. La constante de ponderación es 4.685.

c. Las constantes de ponderación son 1.700, 3.400 y 8.500.

d. La constante de ponderación es $1.340 \cdot \pi$.

Valores extremos

		Número del caso		Valor
Situación del Entrevistado en el Futuro	Mayores	1	995	10
		2	1101	10
		3	586	10
		4	41	10
		5	564	. ^a
	Menores	1	114	0
		2	790	0
		3	83	0
		4	513	0
		5	380	. ^b

a. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 10.

b. En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 0.

VISUALIZACIÓN DE VARIABLES: GRÁFICOS

Los conceptos estadísticos vistos en la sección anterior pueden ser convertidos y traducidos en gráficos. El dicho de que un gráfico vale más que mil palabras se aplica en este contexto diciendo que vale más que mil números. Tanto las distribuciones como las medidas que dan cuenta de sus características pueden expresarse a través de una representación que permita sacar más rápidamente conclusiones y, en consecuencia, también hacer más accesibles al público el significado de los datos obtenidos. Del mismo modo que los datos sufren un proceso de elaboración matemática a fin de entenderlos mejor y más rápidamente, el propósito de las representaciones gráficas de este apartado es la de exponer las distribuciones para que se puedan extraer conclusiones sobre su contenido y forma. Incluso, se podría decir que el mejor método para descubrir la estructura de los datos es a través de su representación gráfica, pues los refleja de forma sencilla, visible y rápida. Del mismo modo que, aunque se pueda formar una idea de cómo están situadas un conjunto de ciudades mirando sus distancias kilométricas, el mejor método para saber exactamente dónde están es mediante un mapa o representación gráfica.

Los gráficos que se van a presentar en este capítulo son aquellos que sirven para representar la distribución de una variable. Se trata de ver de qué forma los conceptos vistos anteriormente pueden ser mostrados gráficamente. Para mayor claridad, se divide esta exposición en dos secciones: una en la que se presentan los gráficos para variables discretas (nominales, ordinales y cuantitativas con pocos valores) y otra en la que se contemplan las diversas representaciones adecuadas para las variables continuas (variables cuantitativas con muchos valores).

Sin embargo, antes de presentar los distintos tipos de gráficos conviene recordar los elementos comunes que deben aparecer en un gráfico:

a) El título: aunque la mera vista de un buen gráfico no necesite de más explicaciones, siempre es recomendable un buen título para su representación. En primer lugar, porque de esa manera, el que expone el gráfico comunica al lector qué es lo que desea representar y, en segundo lugar, porque en las buenas publicaciones siempre ha de haber un índice de gráficos y la mejor manera de referenciarlo es mediante su título.

b) El subtítulo: generalmente es una segunda línea añadida al título en un tamaño de letra menor, que complementa la información proporcionada por el título.

c) La leyenda ayuda a descifrar los símbolos del gráfico y consiste en un rectángulo donde se detalla brevemente lo que significa cada uno de los símbolos explicados en el párrafo anterior.

d) Etiquetas: son textos aclaratorios que acompañan a aspectos esenciales del gráfico. Pueden ser textuales y, de este modo, identifica al objeto que se marca. O bien, son numéricas en cuyo caso aclaran el valor que está representando un elemento gráfico.

e) Nota a pie: es una anotación voluntaria en el gráfico que se realiza para precisar o resaltar alguna de las características peculiares de los datos,

f) Símbolos del gráfico: son las formas, los colores y las tramas. Las primeras modulan aspectos propios de cada gráfico como pueden ser las barras, los puntos, las líneas, etc.; los segundos también marcan diferencias en los elementos gráficos y pueden funcionar de modo alternativo o complementario con las tramas, o modo de expresión de superficies (barras verticales, horizontales, lunares, líneas onduladas, ...)

g) Ejes: son escalas en las que se ubican las variables representadas. En un gráfico es común tener dos o tres ejes; aunque también los hay con uno o ninguno.

h) Serie: cada una de las variables que son representadas en el gráfico.

i) Marcos: son líneas que envuelven el conjunto o determinados aspectos del gráfico con el propósito de remarcarlos.

j) Finalmente, en todo gráfico pueden añadirse imágenes, textos o elementos geométricos que complementen la imagen formada por los elementos acabados de reseñar.

Gráficos para variables discretas

Ya han quedado definidas las variables discretas como aquellas que poseen un conjunto finito de valores, entre dos cualesquiera de ellos no siempre es posible encontrar un tercero. Por tanto, son variables entre cuyos valores no existe continuidad, sino que hay un espacio o terreno que no pertenece a ninguna de las modalidades.

Los dos gráficos más apropiados para la representación de este tipo de variables son el gráfico de sectores y los diagramas de barras.

El primero consiste en un círculo segmentado en sectores de tamaño proporcional a la frecuencia (absoluta o relativa, pues en ambos casos daría el mismo resultado) de cada uno de los valores de la variable.

Para obtener el tamaño de cada sector, basta con transformar las proporciones correspondientes a cada valor en número de grados del círculo. Como es sabido, un círculo, independientemente de su dimensión, siempre contiene 360 grados, por lo que para la construcción del gráfico de sectores hay que repartir esos grados de modo proporcional al peso que tienen los valores de la variable. Si llamamos g_i al tamaño en grados de cada sector, entonces

$$g_i = 360p_i$$

$$\sum_{i=1}^I g_i = 360 \quad (34)$$

Los gráficos de sectores han de emplearse cuando se tienen variables discretas cuyos valores, preferiblemente pocos -no más de siete- sean mutuamente excluyentes. La razón de que se pidan pocos valores es porque de otro modo no se concentraría la atención en ninguna de las categorías. Además son tanto más elegantes y presentables, cuanto menos categorías presenten. El que no deba emplearse con categorías que no sean mutuamente excluyentes se debe a que al estar cerrados y limitados los distintos valores representados en sectores circulares, provocan la sensación de que son partes independientes las una de las otras que complementan y conforman el todo del círculo, que representa a la muestra o la población. En consecuencia, este gráfico debe emplearse cuando se dispone de un conjunto escaso de categorías que conforman un todo para comparar relativamente su magnitud.

Supóngase que se clasifican los 44 países europeos existentes en 1999 en cinco zonas geopolíticas conforme a la tabla 4.

Tabla 4.- Clasificación geográfica de los 44 países europeos (1999).

Países europeos				
<u>Anglosajones</u>	<u>Centro-europeos</u>	<u>Sur de Europa</u>	<u>Este de Europa</u>	<u>Nórdicos</u>
Irlanda	Alemania	Chipre	Albania	Dinamarca
Reino Unido	Andorra	España	Armenia	Finlandia
	Austria	Grecia	Bielorrusia	Islandia
	Bélgica	Italia	Bosnia-Herzegovina	Noruega
	Francia	Malta	Bulgaria	Suecia
	Luxemburgo	Portugal	Croacia	
	Mónaco	Turquía	Eslovaquia	
	Países Bajos		Eslovenia	
	República Checa		Estonia	
	Suiza		Georgia	
			Hungría	
			Letonia	
			Lituania	
			Macedonia	
			Moldavia	
			Polonia	
			Rumania	
			Rusia	
			Ucrania	
			Yugoslavia	

A pesar de la peculiar disposición de los datos en la tabla precedente, es evidente que en ella está reflejada una variable nominal (*región geopolítica*) con 44 casos y cinco valores mutuamente excluyentes. El primer valor corresponde a los países anglosajones y tiene una frecuencia de dos, el segundo valor responde a diez países del centro europeo y así sucesivamente. Con el número de países que hay en cada una de las zonas se forma un sector en el círculo, como se refleja en la figura 11.

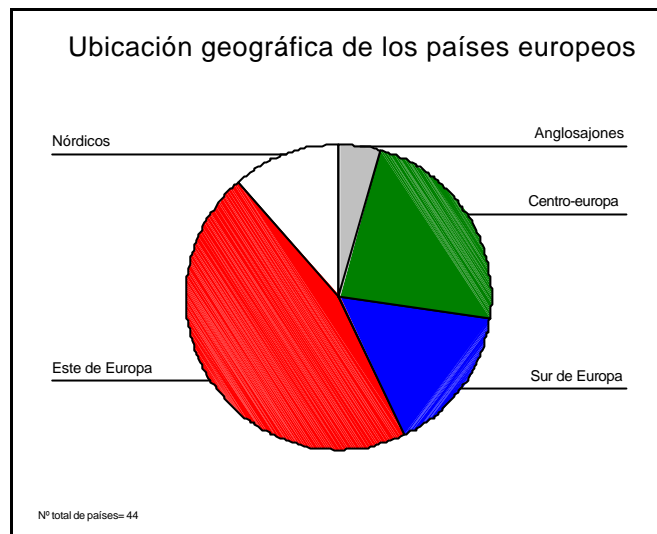


Figura 11.- Gráfico de sectores.

Para obtener este gráfico con el programa SPSS a partir de los datos brutos de una matriz se utilizan las siguientes instrucciones: *Gráficos/Sectores/Resúmenes para grupos de casos*. Tras ello, aparece el siguiente cuadro de diálogo:

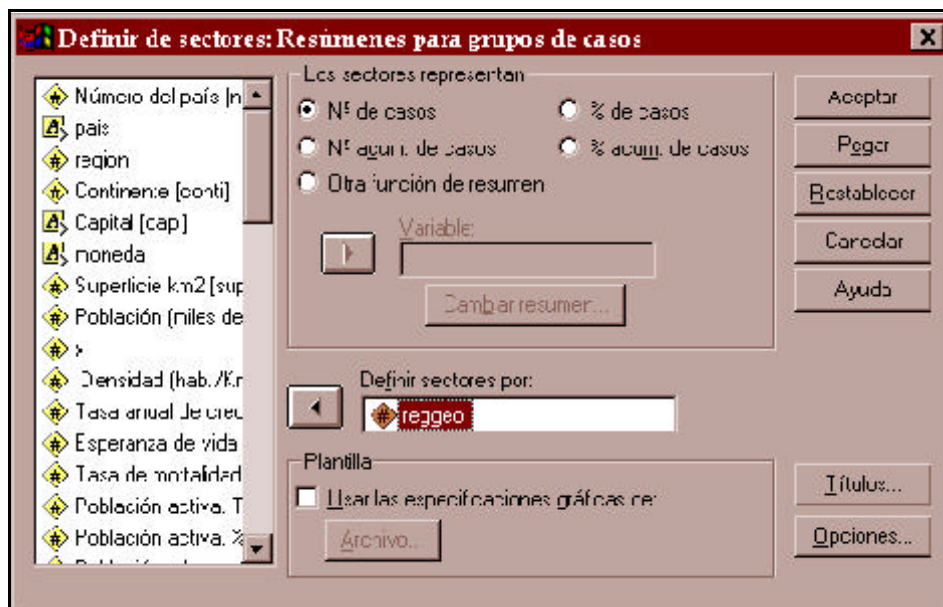



Figura 12.- Cuadro de diálogo para los gráficos de sectores.

Este cuadro de diálogo está compuesto por cinco partes. La primera (izquierda) contiene una ventana con todas las variables del fichero de datos. La segunda (centro superior) permite especificar qué quiere que se represente en los sectores del gráfico. Contiene cuatro botones alternativos que indican si se desea en el gráfico frecuencias absolutas o relativas, simples o acumuladas, y una ventana por si se quiere un gráfico de barras con una cantidad distinta de la frecuencia. En la tercera zona (medio) existe una ventana, denominada *Definir sectores por*, donde se ubica la variable que se ha de representar. La cuarta (centro inferior) permite dar al gráfico un formato distinto del típico del programa. Y en el quinto sector del menú aparecen una serie de botones, que van desde el *Aceptar* hasta el de *Opciones*, donde puede configurarse si el gráfico incluirá o no una categoría con los datos perdidos.

En cambio, si se dispone de los datos ya resumidos, el modo de plantear las instrucciones es algo distinto del acabado de explicar. Supóngase que se dispone de una matriz de datos del siguiente modo:



	zona	países	val
1	Angloesajones	2	
2	Centro-europeos	10	
3	Sur de Europa	7	
4	Este de Europa	20	
5	Nórdicos	5	
6			

Figura 13.- Matriz de datos agrupados.

En cuyo caso, las instrucciones que han de utilizarse son: *Gráficos/ Sectores/Valores individuales de los casos*, tras lo que aparece el menú presente en la figura 14. En su parte central aparecen dos nuevos recipientes de información. El primero se indica bajo la denominación *Los sectores representan*, donde hay que trasladar la variable (países) que indica las frecuencias de los valores que han de ser representadas. Téngase en cuenta que según este modo de construcción del gráfico, el número de casos -la cantidad de filas de la matriz de datos- no es sino el número de valores de la variable (Zona) que se desea representar. El segundo aparece denominado como *Etiquetas de los sectores*. Este bloque permite dos opciones: una que la etiqueta del sector sea meramente el número del caso y otra, la más frecuente, la utilización de los valores -o sus etiquetas- de una variable. En este ejemplo, dado que se tiene una variable alfanumérica (Zona) cuyos valores son textuales, basta con incluirla en la caja de la variable,

después de trasladar la selección desde el *Número del caso* hasta el botón correspondiente a *Variable*.

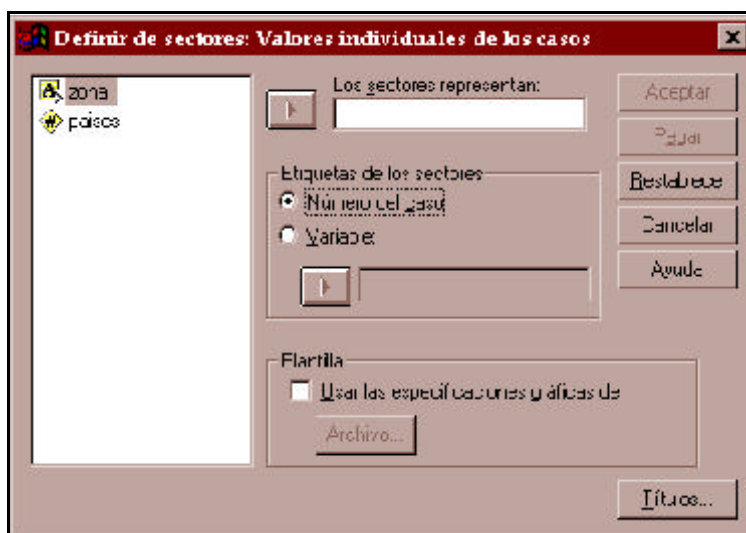


Figura 14.- Menú para gráficos de sectores de valores individuales.

Los gráficos de sectores para una sola variable poseen pocas modalidades de cambio. Además de los colores y tramas de los segmentos del círculo, se puede especificar las etiquetas de los valores sea tanto dentro de los sectores, como fuera de ellos o incluso en el cuadro de la leyenda. Y, desde el punto de vista de la forma, existen dos modalidades distintas: una que implica separar del sector uno - o varios- de los sectores dibujados con el fin de resaltarlos y otra la de dar al círculo un aspecto tridimensional.

Para hacer cambios al gráfico, hay que editarlo después de producido. Para poder realizar las modificaciones, se hace doble clic al gráfico marcado en el visor de resultados, tras lo cual aparece una pantalla como la de la figura 15, que, de hecho, introduce el gráfico en un subprograma con diferentes menús del principal. Las características cambiables en este estadio del gráfico son: la trama (*Formato/Trama de relleno*), el color (*Formato/Color*) y la posición (*Formato/Desgajar sector*) de los sectores; el contenido, las fuentes y el tamaño de los textos (títulos, etiquetas y pie) mediante las instrucciones *Diseño/Título*, *Diseño/Nota a pie de página*, *Diseño/Opciones* y *Formato/texto*; los marcos adicionales del gráfico (*Diseño/Marco exterior*); las categorías de la variable que se quieren mostrar (*Series/Visualizadas*), y, finalmente, se puede cambiar el tipo de gráfico mediante el menú *Galería*. En la mayor parte de los casos, para poder realizar los cambios, es preciso previamente señalar el elemento que se desea modificar. Así en la figura 15, lo que aparece marcado es el texto del título, de modo que si se cambia la fuente, sólo le afectaría a éste y no a otros textos del gráfico.

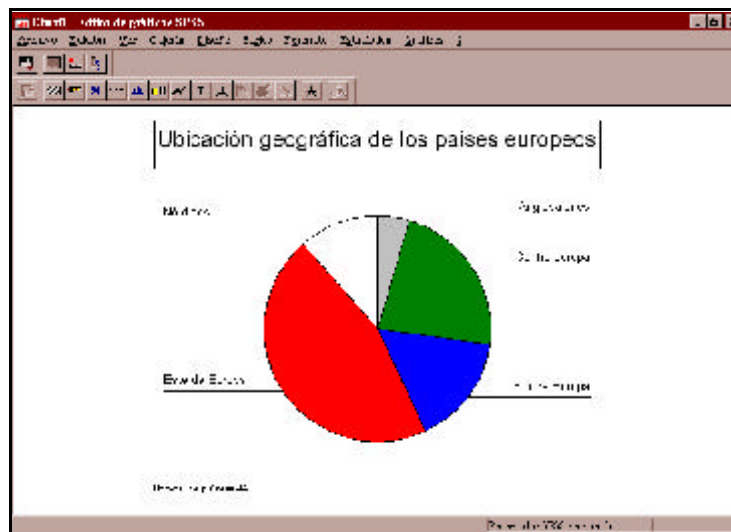


Figura 15.- Pantalla del editor de gráficos.

Otra posibilidad existente desde la versión 8.0 del SPSS es la de crear el gráfico de sectores con la opción interactiva. Tras ordenar mediante menú *Gráficos/Interactivos/Sectores*, aparece el siguiente cuadro de órdenes:

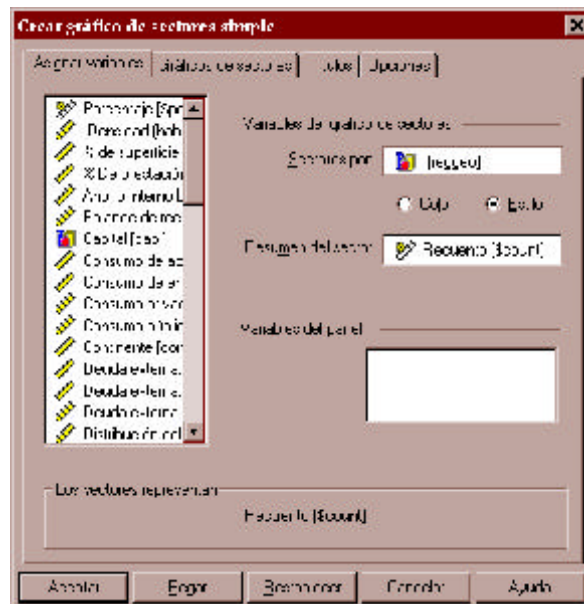


Figura 16.- Cuadro de diálogo para los gráficos de sectores interactivos (*Variables*).

En este cuadro de diálogo, hay que especificar al menos cuál es el *Resumen del sector* (recuento o porcentaje) y la variable (nominal u ordinal) con la que se construye (*Sectores por*). Además en el cuadro aparecen otras subcarpetas (*Gráficos de sectores, Títulos y Opciones*).

Entre ellas la más importante es la primera pues a través de ella se puede especificar qué etiquetas van a constar en el gráfico (las de *Categorías*, las del *Valor* o las que expresan el *Recuento* o el *Porcentaje* que representa cada sector. También se puede concretar si se desea que las categorías se dibujen en la dirección de las manecillas del reloj o a la inversa, así como el grado del círculo donde se comienza a dibujar el primer valor de la variable (figura 17).



Figura 17.- Cuadro de diálogo para los gráficos de sectores interactivos (*Sectores*)

De este modo, podrían obtenerse gráficos más transformados que los anteriores:

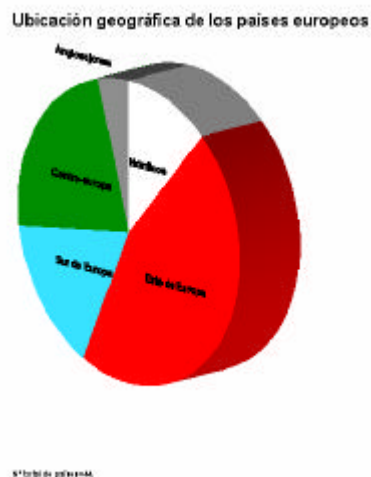


Figura 18.- Gráfico de sectores en tres dimensiones.

El otro tipo de gráfico que se emplea para variables nominales, ordinales y cuantitativas discretas es el diagrama de barras. Consiste en un eje de coordenadas en el que se colocan los distintos valores de la variable en el eje horizontal con un rectángulo cada uno de ellos de altura proporcional a su correspondiente frecuencia (absoluta o relativa).

En consecuencia, el eje horizontal se divide en tantos fragmentos como valores tenga la variable que se desea representar gráficamente levantando desde cada uno de ellos un rectángulo cuya base se diferencia de la(s) colindante(s). Lógicamente, estos gráficos tienen también un eje vertical con una escala que va desde el 0 -frecuencia nula- hasta, al menos, la frecuencia del valor modal.

Como quiera que este tipo de gráficos se emplea para el mismo tipo de variables que usa el gráfico de sectores, se puede poner el mismo ejemplo de variable. No obstante, hay pequeñas diferencias en uso, pues el diagrama de barras permite utilizar sin problema mayor número de valores y, además, no necesariamente requiere que las categorías sean mutuamente excluyentes.

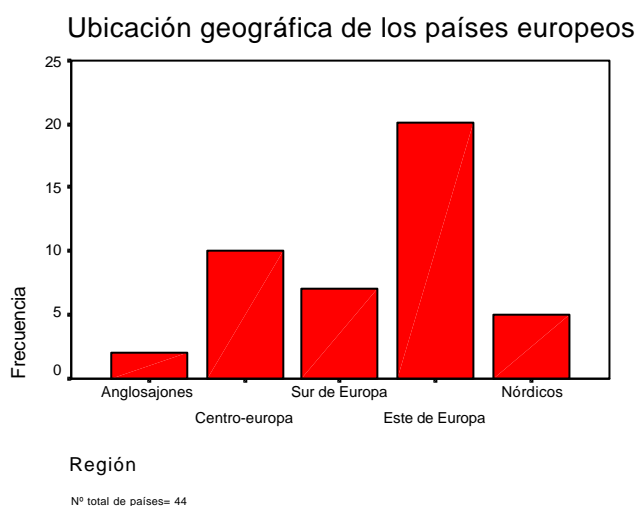


Figura 19.- Diagrama de barras.

Para construir este gráfico, basta con la selección desde el menú de las siguientes órdenes: *Gráficos/Barras/Simple/Resúmenes para grupos de casos*. De esta manera, aparece un cuadro de diálogo similar al de la figura 12, debiéndose pasar la variable que se desea representar a la casilla indicada como *Eje de categorías*.

El diagrama de barras es más susceptible de modificación que el gráfico de sectores. Además de poder cambiar los colores y tramas de los rectángulos; el contenido, las fuentes y el tamaño de los textos, y las categorías de las variables, el subprograma que edita el gráfico permite:

a) Cambiar la orientación del gráfico de modo que las categorías estén representadas en el eje vertical y sus frecuencias en el horizontal. (*Formato/Intercambiar ejes*).

b) Transformar las dimensiones de las barras. Esto se puede lograr tanto cambiando el margen que la separa de los ejes del gráfico, como aumentando o reduciendo la distancia entre ellas. (*Diseño/Espaciado de barras*).

c) Modificar aspectos relevantes de los ejes. (*Diseño/Ejes*). En este tipo de gráficos los ejes son de distinta naturaleza: en el de las categorías se exponen valores no numéricos y lo que realmente importa es la forma de disponer las etiquetas que deben figurar para reconocer los atributos de la variable que se intenta representar. En cambio, es más susceptible de cambio el eje donde figura las frecuencias (*Escala*). En éste (véase figura 20), se pueden modificar el rango (*mínimo y máximo*) de la escala, el modo de la escala (*lineal o logarítmico*), la cuantía y forma de las divisiones y subdivisiones de la escala, y el valor de la escala desde donde se levantan las distintas barras (*Línea de origen de las barras*).



Figura 20.- Cuadro de diálogo para el cambio de los ejes.

d) Dibujar líneas de referencia tanto entre las categorías (verticales) como en la escala (horizontales) con el objeto de mejorar las comparaciones o averiguar más fácilmente la frecuencia de una barra. (*Diseño/Línea de referencia*).

e) Modificar la apariencia de las barras. (*Formato/Estilo de Barras*) Hay en este programa tres estilos diferentes: *Simple*, *sombreado*, y *Efecto 3D*.

f) Insertar texto adicional vinculado a las barras, que puede ser la expresión numérica de sus frecuencias (*Formato/Estilo de formato de barras*) o texto libre a elección del productor del gráfico (*Diseño/Anotación*).

Existe una variante pequeña del diagrama de barras. Se trata del gráfico de Pareto. La primera diferencia con el diagrama de barras estriba en que se ordenan las categorías o valores de las variables según su frecuencia. Así la primera categoría del nuevo gráfico será la de los países del Este, que es la que contiene mayor número de casos. De esta forma, se da mayor

importancia a los valores de la variable que son más frecuentes. La segunda diferencia es la de que se dibuja una línea adicional con la frecuencia relativa acumulada, que indica la concentración de las frecuencias en los valores. Cuanto más horizontal sea la recta, mayor diferencia entre las frecuencias de los distintos valores y, viceversa, a mayor inclinación mejor equilibrio entre las categorías.

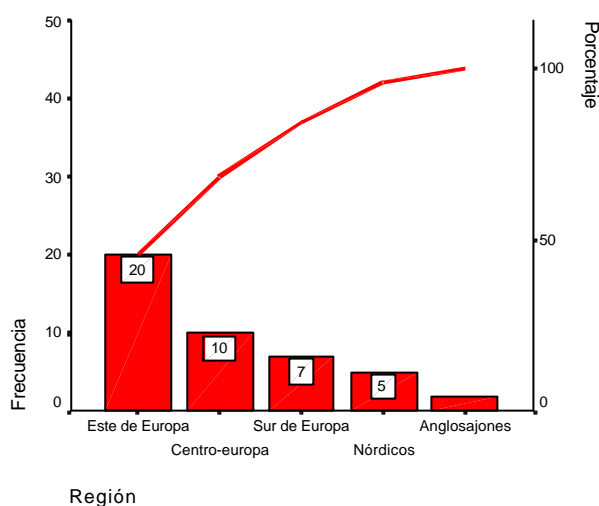


Figura 21.- Gráfico de Pareto.

Gráficos para variables continuas

El primer gráfico de variable continua que se presenta en esta sección es una extensión del diagrama de barras. Así como las barras suelen exponerse separadas, los histogramas dibujan los rectángulos unidos entre sí, indicando de este modo que existe continuidad en la escala de los valores de las variables. Un histograma es, por tanto, un gráfico de variable continua dividida en intervalos de los que se eleva un rectángulo con área proporcional a su frecuencia absoluta o relativa. Pero, como la mayor parte de las veces -y el programa SPSS no lo permite hacer de otra forma- los intervalos son de igual amplitud; en esos casos también podría decirse que la altura de los rectángulos es proporcional a la frecuencia a la que representan.

La construcción de un histograma se realiza con la instrucción *Gráficos/Histograma*, tras lo cual aparece un sencillo cuadro de diálogo, en el que sólo se puede indicar la variable cuyo gráfico se desea obtener, los títulos que se quieren poner, la plantilla que se desea utilizar y la

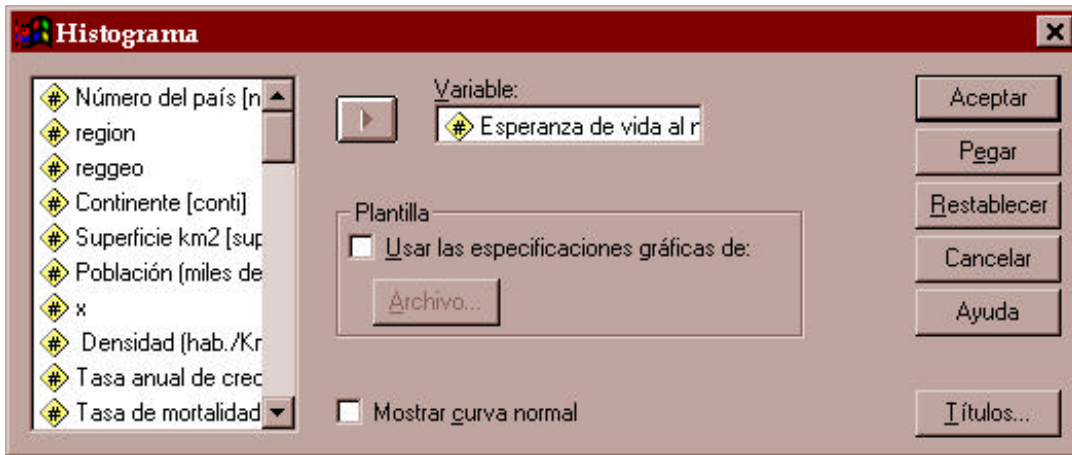


Figura 22.- Cuadro de diálogo para histogramas.

inclusión en el gráfico de la curva normal.

Con tan pocos parámetros permitidos, la confección del gráfico es incontrolable pues automáticamente se calcula, a partir de los valores máximos y mínimos de las variables, el número y la amplitud de los intervalos. Esto queda patente en la figura donde se representa la esperanza de vida al nacer de 41 países europeos en ocho intervalos con un tamaño de un par de años. Así en el primer intervalo quedan incluidas los valores 65 y 66, en el segundo los de 67 y 68 y, así sucesivamente hasta llegar al último intervalo donde se representan los países con esperanza de vida al nacer de 79 y 80 años.

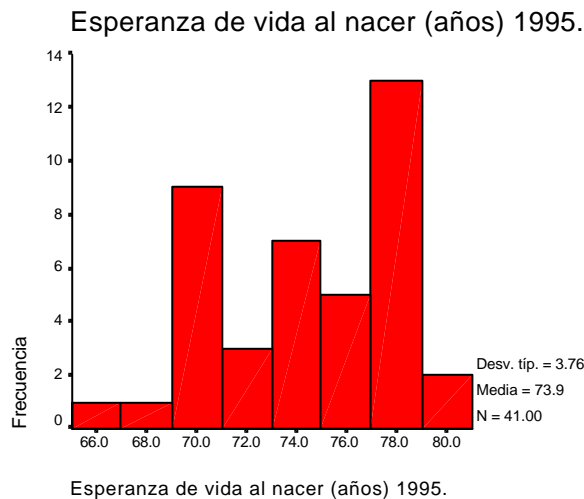


Figura 23.- Ejemplo de histograma.

En realidad, lo que se ha efectuado es una recodificación automática por el programa que luego es representada como si fuera un diagrama de barras con la única limitación de que no puedan separarse los rectángulos. Por ello, las modificaciones ejecutables sobre este gráfico, con la excepción acabada de mencionar, son idénticas a las comentadas en la página 51. Otra

diferencia importante con respecto al diagrama de barras es que calcula automáticamente una leyenda donde consta el número de casos, la media y la desviación típica.

En el gráfico de las esperanzas de vida de los países europeos se comprueba que la moda está en el intervalo que incluye los 77 y los 78 años, esperanza en la que se encuentran casi un tercio de los casos analizados, aunque también haya un importante número (más del 20% de naciones con esperanza entre los 69 y los 70 años).

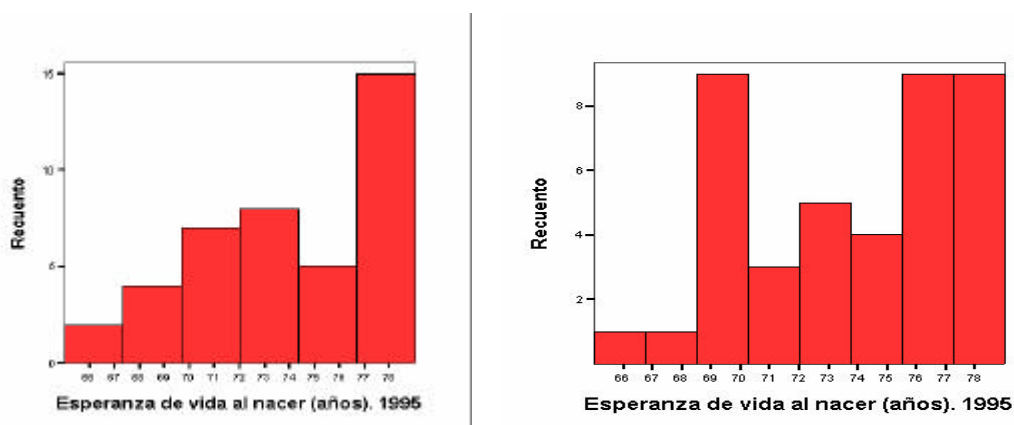


Figura 24.- Comparación de histogramas con distinto número de intervalos.

Mayor posibilidad de cambio proporciona la herramienta *Gráficos/Interactivos/Histograma*, que permite especificar el número de intervalos que se requiere representar. El inconveniente, en este caso, es que lo automatiza de tal modo que los puntos de corte son números decimales sin ninguna regularidad. Cuando se trata de variables continuas empíricamente, esto apenas produce desajustes; pero si, en realidad, se están representando variables discretas, la interpretabilidad de los histogramas es muy deficiente. Véase, como prueba de ello dos gráficos que se han obtenido con la opción interactiva, una vez con 6 intervalos y la otra con 8 y compárese con la proporcionada en la figura 23. Incluso con el mismo número de intervalos las divisiones no son las mismas, pues en la posterior no se centran los intervalos en el punto medio. Aunque no parezca cierto, los gráficos de las figuras 23 y 24 proceden de los mismos datos.

Los gráficos de línea o, también llamados, polígonos de frecuencias consisten en unir los puntos medios de todos los histogramas contiguos mediante una recta. Sin embargo, el programa SPSS (*Gráficos/Líneas*) no lo realiza de esta forma, sino que opera como si estos gráficos fueran de variable discreta y lo que efectúa es la unión de rectángulos de los diagramas de barra, en lugar de hacerlo con los histogramas. Este no es problema si todos los valores reales de la variable son equidistantes; pero en el momento que no lo sean se plantean problemas de inexactitud gráfica. El mismo ejemplo que ha servido para dibujar los histogramas ayuda a comprender las diferencias entre ambas formas de representación, porque existe en la distribución dos valores intermedios sin ninguna frecuencia (66 y 68 años):

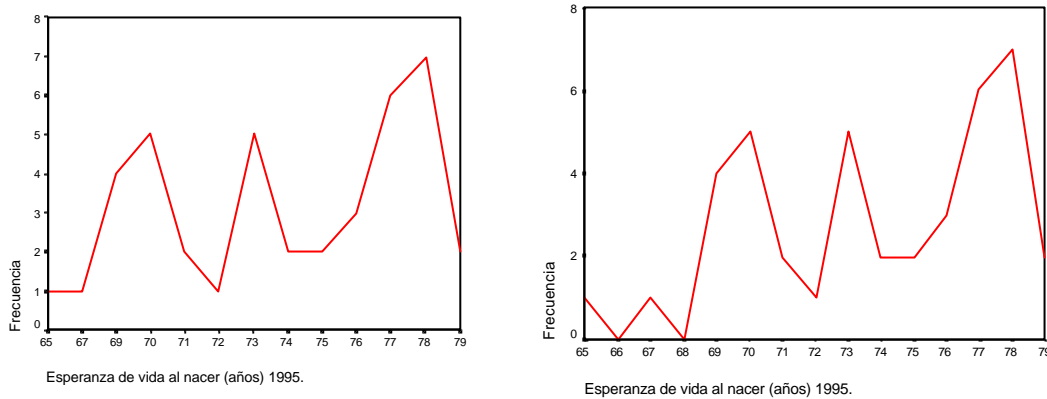


Figura 25.- Gráficos de líneas representados como variable discreta y continua.

Gráficos muy similares a los de líneas son los gráficos de área. Lo único que los diferencia es que la zona existente entre el polígono de frecuencias y los ejes de coordenadas se rellena con un color con el fin de subrayar la presencia de casos en el interior.

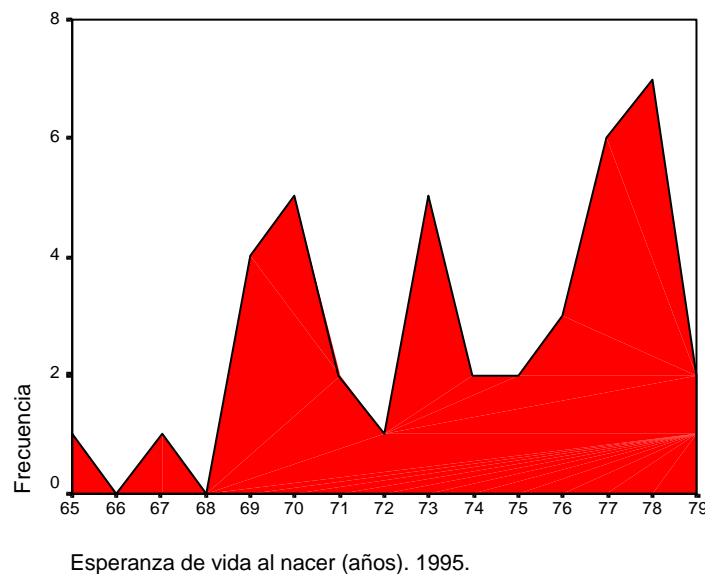


Figura 26.- Ejemplo de gráfico de área.

Los gráficos de tallo y hoja son un modo fácil y evidente de representar semigráficamente una variable cuyos valores sean números. Como su propio nombre indica son representaciones de datos que constan de dos elementos: el tallo y las hojas. La primera cuestión que debe resolverse para abordar estos gráficos es qué es lo que se va a considerar tallo y qué se entenderá como hoja.

El tallo está constituido por el/los primeros dígitos de la variable. La hoja sería el siguiente dígito no incluido en el tallo. La dimensión de este último está determinada por las características

del rango de la variable. Se trata de formar entre 5 y veinte filas de números. Como no hay normas ni fórmulas precisas, lo mejor es dar algunos ejemplos de aplicación.

Caso de disponer de una variable como la edad con los siguientes valores

{3, 5, 7, 8, 10, 14, 17, 20, 25, 30, 32, 35, 37, 39, 41, 50, 53, 60, 72, 81}

es conveniente considerar como hoja (derecha) el último dígito (unidades) y como tallo (izquierda) el primer dígito (decenas). De este modo, la distribución quedaría como sigue:

0	-357
1	-047
2	-05
3	-0257
4	-1
5	-03
6	-0
7	-2
8	-1

Figura 27.- Gráfico de tallo y hoja (edad).

En este gráfico un tallo equivale a 10 unidades y cada una de las hojas representa una unidad. Así, es fácil deducir que en la muestra considerada hay cuatro personas con menos de diez años, tres comprendidas entre los 10 y los 19, cuatro entre los 30 y los 39, y así sucesivamente hasta ver que sólo hay una persona con más de ochenta años.

En cambio, para representar la esperanza de vida de los países europeos, si se procediera con la misma estrategia que en el anterior caso, sólo habría 2 filas: una muy corta con sólo 6 casos, los seis países con una esperanza de vida inferior a 70 años y 35 en la segunda. En lugar de ello, se toma como determinación que el ancho del tallo equivalga a 10 unidades; pero se representa dos de ellas en cada una de las líneas. De este modo el gráfico quedaría configurado como sigue:

Esperanza de vida al nacer (años) 1995.

Frecuencia	Tallo & Hoja
1.00	6 . 5
1.00	6 . 7
4.00	6 . 9999
7.00	7 . 0000011
6.00	7 . 233333
4.00	7 . 4455
9.00	7 . 666777777
9.00	7 . 888888899
Ancho del tallo: 10.00	
Cada hoja: 1 caso	

Figura 28.- Gráfico de tallo y hoja (esperanza de vida).

Otro ejemplo es el del producto nacional bruto de los países europeos cuyo rango va desde los 2.252.343 millones de dólares del PNB alemán hasta los 800 que produce Mónaco. Como en estos datos, hay cinco países que se desvían enormemente del conjunto (Alemania, Francia, Reino Unido, Italia y España), se crea una categoría especial para ellos denominada extremos y el resto queda dividido en líneas de 50.000 millones de dólares. Para su representación el dígito del tallo equivale a las centenas de millar de millón, mientras que las hojas, una por país, indican las decenas de millares de millón.

PNB. Millones de dólares. 1995.

Frecuencia	Tallo & Hoja
24.00	0 . 000000000000000111112334
4.00	0 . 5889
3.00	1 . 003
2.00	1 . 56
2.00	2 . 01
2.00	2 . 58
1.00	3 . 3
1.00	3 . 7
5.00	Extremes (>=532347)
Ancho del tallo: 100000.0	
Cada hoja: 1 caso	

Figura 29.- Gráfico de tallo y hojas (PNB).

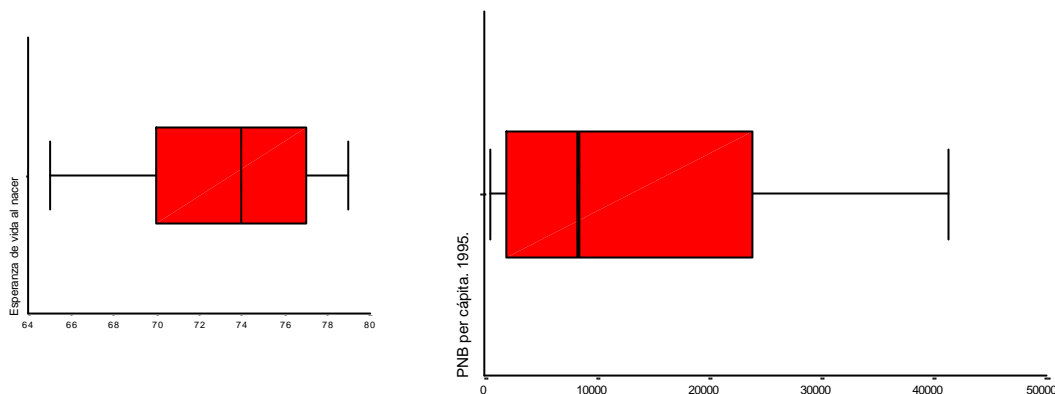


Figura 31.- Diagramas de caja de las variables esperanza de vida al nacer y PNB per cápita.

Estos dos gráficos muestran respectivamente una distribución asimétrica a la izquierda en el caso de la esperanza de vida ($As = -0.427$) y a la derecha en el del PNB per cápita ($As = 0.812$). La mediana de la esperanza de vida está situada en el valor 74, los cuartiles primero y tercero en 70 y 77 respectivamente. El valor mínimo (65) está representado por la línea vertical de la izquierda y el valor máximo (79) también está fijado por la línea vertical de la derecha. En ambos casos, como también en la figura correspondiente al PNB, las rectas que salen de la caja llegan hasta los casos mínimo y máximo porque ninguno de ellos está situado a una distancia de su cuartil más próximo punto y medio superior al rango intercuartílico. En el caso de la esperanza de vida el valor mínimo es 65 y el primer cuartil es 70, cuando el rango intercuartílico es de 7 puntos. A su vez, el valor máximo en el PNB per cápita es de 41.000 dólares, el tercer cuartil tiene un valor de 24.000 y el rango intercuartílico es incluso superior (22.000) a la distancia entre las dos anteriores cantidades.

A aquellos casos que se ubican por encima de un rango y medio intercuartílico del tercer cuartil o por debajo de menos un rango y medio el primer cuartil se les denomina atípicos y, entre ellos, reciben el nombre de extremos los que se alejan de su cuartil más próximo tres veces el rango intercuartílico.

Así, en el caso de representar el PNB global, en lugar del per cápita, se encuentran un país atípico (España) y cuatro extremos (Alemania, Francia, Reino Unido e Italia) como se puede apreciar en la figura 32.

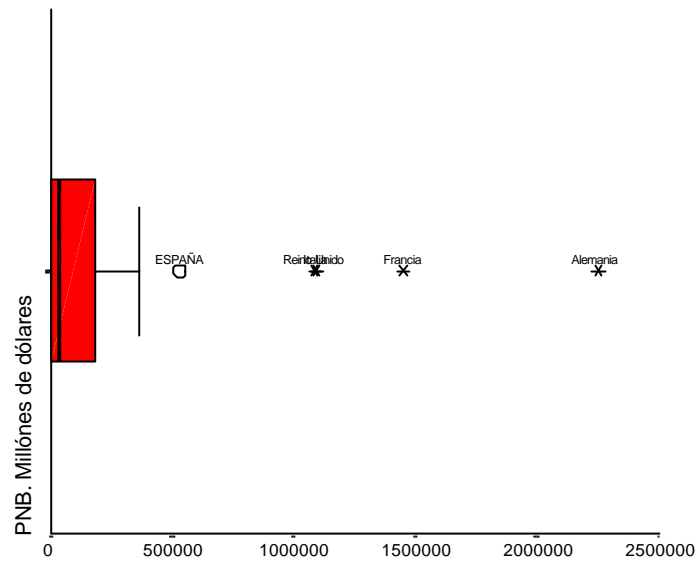


Figura 32.- Diagrama de caja del PNB de países europeos.

Para obtener estos gráficos univariados con el programa SPSS a través de menú, las instrucciones son *Gráficos/Diagramas de caja/Simple/Resúmenes para distintas variables*. Como sus antecesores, con la diferencia del gráfico de tallo y hoja, pueden ser editados en una pantalla específica para su transformación. Los cambios que pueden hacerse afectan a textos (títulos, pie, anotaciones, leyenda y valores y etiquetas de ejes), las líneas (marcos, ejes, líneas de referencias), la escala, la caja, y los casos atípicos y extremos.

TRANSFORMACIÓN DE VARIABLES

Cuando se trabaja con variables, hay que tener en cuenta que en muchas ocasiones no es cómodo o adecuado a los propósitos del análisis utilizar las mismas unidades o categorías con la que fueron recogidas o, con la que fueron introducidas en la matriz de datos informatizada. Un ejemplo claro puede ser la edad. Imagínese que se tiene un fichero con una variable consistente en el año de nacimiento; pero lo que interesa no es cuándo se nació, sino la edad que tienen los sujetos. Para ello, puede procederse a transformar la variable año de nacimiento en otra denominada edad mediante la sustracción de la primera al año en el que se llevó a cabo la encuesta. O, incluso, piénsese en una escala de Likert con varios ítems con valores de 1 a 5, en los que unos van en una dirección de la actitud (valores ecologistas, por ejemplo) y otros en la dirección contraria. A fin de obtener una medida aditiva, es obvio que las variables correspondiente a los ítems expresados en sentido inverso al que se desea medir necesitan ser transformadas, convirtiendo los 5 a 1, los 4 a 2, los 2 a 4 y los 1 a 5. Otras veces, incluso, se necesitan cambios de escala más intensos, que afectan a la distancia entre los puntos de la escala. Así, es práctica frecuente convertir variables numéricas, sobre todos rentas e ingresos, a escala logarítmica. Ello produce cambios importantes en la naturaleza de la variable que se examinarán más adelante. Finalmente, hay otras veces que se necesita cambiar el tipo de variable; por ejemplo, convertir una variable nominal en otra de intervalo con el fin de realizar con ella un análisis para el que se precisa este nivel de medición, o, en otras ocasiones, no conviene exponer todos los valores de la variable y procede la agrupación de varios de ellos, siendo una de las posibles razones el que haya pocos casos con un determinado valor de la variable.

En resumen, los cambios sobre los valores de las variables se clasifican en tres tipos. En primer lugar, se distinguen las transformaciones aritméticas de las que son lógicas. Y, entre las transformaciones aritméticas se distinguirán dos tipos con propiedades muy distintas entre sí: las transformaciones lineales y las no lineales.

Una transformación es lineal si puede ser reducida a la siguiente expresión:

$$x_i' = a + bx_i \quad (35)$$

donde x_i son los valores de la variable original, x_i' los de la variable transformada y a , b parámetros constantes que poseen un valor u otro en función del cambio que se desee realizar.

Si, por ejemplo, se desea transformar la fecha de nacimiento en edad, los valores que han de adoptar los parámetros a y b son, respectivamente, el año de realización del estudio y -1 . Así, caso de haberse realizado la encuesta en el año 1998, obtendríamos la edad de los entrevistados mediante la siguiente expresión:

$$x_i' = 1998 - x_i \quad (36)$$

Otro ejemplo útil es el de la conversión de los items invertidos de la escala de Likert. En estos casos, el valor de b siempre será -1 y el valor de a la suma de los valores máximo y mínimo de la escala. Así, para convertir una variable de rango 1 a 5 a otra de rango inverso de 5 a 1, la fórmula que tendría que aplicarse sería:

$$x_i' = 6 - x_i \quad (37)$$

Y, en el caso de desear convertir una escala de 0 a 10, en otra de 10 a 0, la ecuación se transformaría en:

$$x_i' = 10 - x_i \quad (38)$$

Las transformaciones lineales tienen una serie de propiedades importantes que las diferencian del resto de transformaciones. Estas son:

$$\begin{aligned} \bar{x}_i' &= a + b\bar{x} \\ \text{Var}(x_i') &= b^2 \text{Var}(x) \end{aligned} \quad (39)$$

Es decir, la media de una variable transformada puede obtenerse mediante la transformación lineal de la variable original; sin embargo, la varianza de la nueva variable se obtiene mediante la multiplicación de la antigua varianza por el cuadrado del parámetro b. Esto conlleva, por ejemplo, en el caso de las transformaciones vistas en las ecuaciones (36) a (38) que la varianza no cambie a pesar de la mutación de la variable, pues en esos tres casos $b=-1$.

Sin embargo, si una variable expresada en miles de pesetas, se transforma en otra medida en pesetas, el coeficiente b sería igual a 1.000 y, en consecuencia, habría que multiplicar la varianza de la primera por 1.000.000 para obtener la de la segunda. En cambio, para la obtención de la desviación típica bastaría con multiplicar por mil la original.

Es importante decir también que sea cual sea la transformación lineal de una variable, la forma de la distribución no cambia. Así, si se calculan la asimetría y curtosis de dos variables que sean una la transformación lineal de la otra, los resultados son idénticos. A modo de ejemplo empírico, en la tabla 5, se han calculado los estadísticos del PNB per cápita, medido en una columna en dólares, en la otra en miles de dólares. Como puede fácilmente apreciarse, la media en la columna de la izquierda es 1.000 veces superior, la varianza 1.000.000; pero los coeficientes de simetría y apuntamiento (asimetría y carosis) son iguales.

Tabla 5.- Estadísticos del PNB per cápita.

	PNB per cápita. 1995. Dólares	PNB per cápita. 1995. Miles de dólares
N	44	44
Mínimo	440.00	.44
Máximo	41210.00	41.21
Media	12120.4545	12.1205
Varianza	143167567.230	143.168
Asimetría	.812	.812
Curtosis	-.469	-.469

Una de las transformaciones lineales más utilizadas en el análisis de datos es la tipificación de una variable. Consiste en cambiar sus valores para que tenga una media de 0 y una desviación típica de 1. De este modo, se pueden hacer comparaciones entre distribuciones de variables medidas en unidades muy distintas, como por ejemplo, pesetas y años. Y, asimismo, permite la comparación de las variables con modelos teóricos de distribución como la normal.

Los valores que adoptan a y b en el caso de la tipificación son respectivamente $-O/s$ y $1/s$; pero el modo más fácil de realizar la conversión de los valores originales x en los valores tipificados es

$$z_t = \frac{x_t - \bar{x}}{s} \quad (40)$$

Existen otro tipo de transformaciones algebraicas de las variables que no son lineales y que tienen una gran importancia en el análisis exploratorio de datos. Son aquellas en las que se aplican funciones distintas de la suma y la multiplicación de constantes. Estas transformaciones tienen la propiedad de cambiar la distancia absoluta y relativa entre los valores y , por tanto, modifican la forma de la distribución de los datos.

De este modo, una transformación puede dispersar los valores en un determinado lado de la distribución o agrupar observaciones que estén muy distantes del valor medio de la variable. En consecuencia, tiene efectos que las transformaciones lineales son incapaces de provocar, como la reducción de la asimetría de una distribución, el cambio de la carosis e incluso hacer más representativos los valores centrales de una variable.

Es común denominar a estos cambios de variables transformaciones de potencia, pues adoptan, en general, la forma:

$$\begin{aligned}
 x_i' &= x_i^q & q > 0 \\
 x_i' &= \log x_i & q = 0 \\
 x_i' &= -x_i^q & q < 0
 \end{aligned}
 \tag{41}$$

La nueva variable se obtiene elevando a una determinada potencia el valor de la variable original, salvo en el caso en que $q=0$, en el que se transforma la variable x aplicándole su logaritmo (neperiano o decimal).

Estas transformaciones tienen la principal virtud de cambiar la forma de la distribución de modo que si $q < 1$, la distribución gana asimetría a la izquierda y, en el caso de que $q > 1$, al contrario, la distribución se hace más asimétrica a la derecha. Evidentemente, si $q=1$, la variable no sufre ninguna modificación.

Figura 33.- Distribución de X .

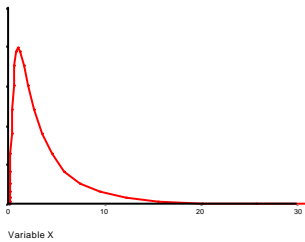


Figura 34.- Distribución de $\ln(X)$.

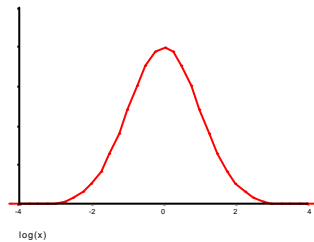
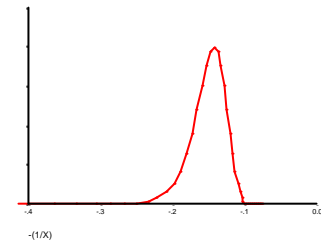


Figura 35.- Distribución de $-(1/X)$.



En las figuras de arriba se ve el efecto que provoca las transformaciones no lineales. Inicialmente, la variable X tiene una distribución asimétrica a la derecha (figura 33). Al aplicarle la función logarítmica, se convierte en una distribución simétrica y normal (figura 34). Pero, si se le hubiese aplicado la función inversa ($1/X$), la simetría habría pasado a ser a la izquierda (figura 35).

A la inversa también es cierta la relación. Si una distribución es asimétrica a la izquierda, tras aplicarle el cuadrado tiende a perder asimetría, tanta más cuanto más se eleve la potencia, llegando incluso a tener una asimetría a la derecha.

Otras transformaciones más complejas operan sobre las colas de la distribución de modo que las suaviza. Se las conoce como logaritmos y raíces plegados. Se ejecutan en dos pasos: en el primero se realiza una transformación lineal de tal naturaleza que la nueva variable tenga como límites los valores entre 0.05 y 0.95. Para ello los valores de los parámetros de la conversión han de ser los siguientes:

$$b = \frac{0.90}{x_{\max} - x_{\min}}$$

$$a = \frac{0.05x_{\max} - 0.95x_{\min}}{x_{\max} - x_{\min}}$$
(42)

Una vez limitados los valores de la variable en el intervalo [0.05-0.95] se le aplica la función que efectivamente realiza la suavización de las colas. En el caso de utilizar logaritmos, la expresión que habría que utilizar es

$$\frac{\ln(x')} {2} - \frac{\ln(1-x')} {2}$$
(43)

mientras que si se emplea las raíces, la transformación que hay que aplicar es

$$\sqrt{(2x')} - \sqrt{(2(1-x'))}$$
(44)

En la tabla 6 se muestra la variable edad con los datos de cuatro personas a la que se le aplica las transformaciones señaladas y, para ver sus efectos, se devuelve a la escala de edad con los mismos valores extremos. La primera columna incluye los valores originales, la segunda su transformación lineal para que el valor mínimo sea 0,05 y el máximo (0,95), en la tercera se aplica el logaritmo plegado sobre el que se vuelve a reescalar para que mínimos y máximos sean otra vez 5 y 35. Como puede observarse, el valor original de 6 se ha convertido en 7.2 y el de 34 en 32.4, con lo que se suaviza la cola de la distribución. Las dos siguientes columnas son paralelas a las dos anteriores; pero con la utilización de las raíces plegadas. En este caso, la diferencia está en que el suavizado de los extremos es menor pues el valor 6 pasa a ser sólo 6.5 y el 34, 33.5.

Tabla 6.- Transformaciones de logaritmo y raíz plegados sobre la edad.

x_t	x_t^I	ln	x_t^{II}	raíz	x_t^{III}
5	5	-147	50	-106	50
6	8	-122	76	-96	65
34	92	122	324	96	335
35	95	147	350	106	350

Es obvio que todas estas transformaciones (lineales y de potencia) sólo se pueden aplicar a variables cuantitativas. Hay otras transformaciones, a las que denominaremos lógicas, que podrían emplearse en todo tipo de variables. Mediante ellas, se transforman con una relación de equivalencia todos o parte de los valores de la variable.

De todas las posibles transformaciones lógicas, la más conocida y empleada es la disposición en intervalos de una variable cuantitativa. Mediante esta operación, uno o varios conjuntos de valores contiguos se funden en un solo valor. Sea, por ejemplo, la edad. Después de entrevistar a 1000 sujetos de edades comprendidas entre los 18 y los 85 años, no resulta cómodo presentar una tabla de distribución de frecuencias con 68 líneas, una para cada edad; sino más bien lo que se realiza es agrupar valores contiguos.

Esta agrupación puede hacerse con criterios muy distintos. Baste señalar los más empleados:

- El primero consistiría es realizar intervalos de tamaño regular. Para ello, se divide el rango entre los intervalos que se desea obtener y con ello se obtiene la amplitud de cada agrupación. Siendo R el rango de la variable e I la amplitud del intervalo, se establece la siguiente relación:

$$a_i = \frac{R}{I} \quad (45)$$

En este caso, como el rango comprende 68 edades distintas, al dividir por cuatro, arroja una amplitud de 17 años. Comenzando por el valor mínimo de 18, los intervalos serían los siguientes: a) 18-34; b) 35-41; c) 42-58; d) 59-85¹⁰.

- El segundo método es el empleo de criterios empíricos. Se trataría de agrupar categorías de modo que los intervalos tengan frecuencias similares. O bien, más frecuentemente, agrupar los valores de aquellas categorías que no tengan un considerable número de sujetos. Un ejemplo muy común es el del número de hijos. Al comienzo de la variable no se hace recodificación: 0, 1, 2, 3, ...; pero a partir de una determinada cantidad hay que realizar agrupaciones (4,5) (6,7),... E, incluso, en los últimos intervalos conviene dejar el último valor abierto (8 o más).

- Habría un tercer criterio de agrupación de valores, que se basa en la agrupación de valores en función del sentido próximo que posean. Así, una agrupación de la variable edad podría realizar el siguiente 18-29 años, 30-39, 40-49, 50-64 y 65 y más. En el primer intervalo se trataría de agrupar a los jóvenes, posteriormente habría dos grupos con una amplitud de una

10 Al tratarse de una variable potencialmente continua, los límites de los intervalos reales deberían coincidir: así, el primer intervalo podría representarse como 17.5-34.5, el segundo como 34.5-41.5, el tercero como 41.5-58.5 y el último 58.5-85.5. En el caso de la edad, sería más conveniente poner como punto de corte los años enteros, pues las personas al declararla trunca a la baja los años que tiene. Desde el punto de vista matemático, pues, la presentación ideal de los intervalos sería [18-35)/(35-42)/(42-59)/(59-85). La marca “[“indica que la edad que le sigue está incluida en el intervalo, el paréntesis, “)” aclara que el valor que le antecede no se incluye en el intervalo.

década, el grupo entre 50 y 64 intentaría reflejar la prejubilación, mientras que a partir de los 65, estaríamos ante la tercera edad.

No toda transformación lógica de los valores de la variable tiene que ser por intervalos contiguos. Existen agrupaciones de valores en las que es preferible que categorías no próximas queden unidas. Piénsese en una lista de partidos políticos a los que la gente tendría intención de votar en unas próximas elecciones: PP, PSOE, CiU, IU, PNV, ...Caso de agrupar estas categorías, habría dos posibilidades: una el eje derecha-izquierda en el que, por un lado, quedarán PP, CiU y PNV, y, por el otro PSOE e IU. Otro modo de agregarlas sería considerando juntos PP, PSOE e IU, partidos centralistas, diferenciándolos de PNV y CiU, que son nacionalistas.

Entre las transformaciones que pueden realizarse sobre las variables nominales, es preciso detenerse en la creación de variables ficticias. Este es un proceso necesario para transformar las variables de tipo cualitativo en otras que tengan también propiedades cuantitativas.

Para ver estas transformaciones, es útil empezar por una variable nominal de dos valores, el sexo, por ejemplo. Piénsese en una muestra con 600 mujeres y 400 varones. Es evidente que no se pueden realizar operaciones aritméticas con los valores “mujer” u “hombre”. En su lugar, se puede hacer un cambio a los valores de modo que uno de ellos esté presentado con el número 1, y el otro con el 0. En ese caso, la variable dejará de ser “exactamente” el *sexo*, para pasar a ser *hombre* o *mujer*. Elíjase ésta última categoría y, de este modo, la distribución de frecuencias quedaría así:

Tabla 7.- Distribución de la variable ficticia *Mujer*

MUJER					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No	400	40.0	40.0	40.0
	Sí	600	60.0	60.0	100.0
Total		1000	100.0	100.0	

Y, si se calculan los estadísticos propios de las variables cuantitativas, se obtienen los siguientes resultados:

Tabla 8.- Estadísticos de la variable ficticia *mujer*.

Estadísticos		
MUJER		
N	Válidos	1000
Media		.600
Varianza		.240

Como puede comprobarse la media coincide con la proporción de casos que tienen el atributo al que arbitrariamente se le ha otorgado el valor de la unidad y la varianza se ajusta a la expresión $p(1-p)$.

Caso de que se tuvieran variables nominales con más de dos valores, las transformaciones en variables ficticias han de repetirse tantas veces como valores menos 1 tenga la variable original. De este modo, para poder representar plena y cuantitativamente una variable nominal de cinco valores, se necesitarían cuatro variables ficticias. Sea, por ejemplo, el estado civil con valores soltero, casado, separado, divorciado y viudo. Los datos de la muestra quedarían reflejados completamente mediante cuatro variables dicotómicas, de valores 0 y 1. Hay, en consecuencia, cinco modos distintos de expresarlas: primera, dejando como base el valor soltero y creando cuatro variables en las que el valor 1 represente respectivamente a las categorías distintas de la de ser célibe; segunda, aislando el valor de casado y conformando cuatro variables binarias con los atributos soltero, separado, divorciado y viudo; tercera, utilizando como variable base de comparación los separados; cuarta, los divorciados, y quinta, los viudos.

Sea cual fuere la variable dejada como base, las cuatro restantes, siempre y cuando se parta de una variable con valores mutuamente excluyentes, sumarán una proporción inferior a 1. El complemento para llegar a la unidad es la proporción de casos que poseen la categoría considerada como base. Volviendo al ejemplo fácil del sexo, si había una proporción de 0.60 mujeres, es fácil deducir que se encontraría en la muestra un 0.40 de hombres. En el caso del estado civil, si la proporción de casados es de 0.60, la de separados de 0.02, la de divorciados 0.03 y la de viudos 0.10, es fácil deducir que la de solteros sería de .25.

Instrucciones para transformar variables

En el programa SPSS hay dos modos de alterar los valores de las variables, que se corresponden con las llamados cambios algebraicos y lógicos. Para el primero se utiliza la orden *Transformar/Calcular*, mientras que para el segundo se emplea la instrucción *Transformar/Recodificar*. Ambas pueden estar restringidas por una condición y realizarse no sobre toda la muestra o conjunto de datos disponibles, sino sólo sobre un subconjunto dado.

La orden *Calcular* genera el cuadro de diálogo expuesto en la figura 36. En él es obligatorio escribir en los dos espacios superiores. En el izquierdo ha de indicarse el nombre que va a tener la variable transformada. Básicamente, hay dos posibilidades, ponerle el mismo nombre que a la variable original, o darle un nombre distinto. Esta última es preferible, pues de este modo, se conserva en la matriz de datos la variable original y se puede hacer con ella nuevas transformaciones. De lo contrario, a menos que se guarde el fichero de datos con otro nombre, se perderán los valores de la variable inicial. En el espacio superior derecho, se ha de escribir la operación algebraica que transforma a la variable.

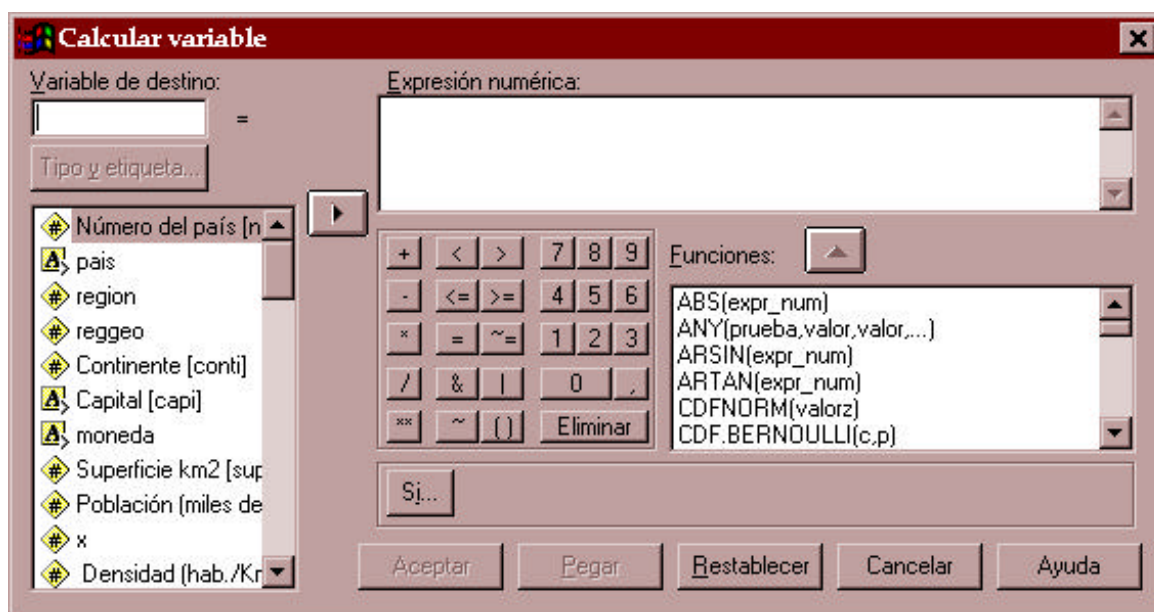


Figura 36.- Cuadro de diálogo para el cálculo de variables.

En los sectores inferiores de la pantalla de la misma figura 36, aparecen, en la izquierda, los nombres de las variables de la matriz de datos; a la derecha, un conjunto de funciones¹¹ que se pueden utilizar para obtener la *Variable de destino*, y en el centro hay una especie de calculadora donde aparecen números por si es preciso utilizar constantes en la *Expresión numérica* y una serie de funciones aritméticas (suma, resta, multiplicación, división y potencia) y lógicas (mayor y menor que, igual y desigual, conectores lógicos (& y |) y negación (~).

Para las transformaciones lineales, basta con escribir en la ventana superior derecha una expresión numérica del tipo:

$$a + b * \text{nombre} \quad (46)$$

En cambio, para las transformaciones de potencia, la expresión adopta la siguiente fórmula:

$$\text{nombre} ** (q) \quad (47)$$

Por ejemplo, para transformar una variable medida en dólares en otra expresada en miles de dólares, se podría utilizar cualquiera de las dos fórmulas indicadas en cada una de las dos líneas siguientes:

$$0 + (1/1000) * PNB \quad (48)$$

$$PNB / 1000$$

¹¹ Las funciones que dispone el SPSS para la transformación de variables son de diverso tipo: aritméticas, estadísticas, generadoras de distribuciones, y funciones de valores perdidos, de retardo, de fecha y de cadena.. En este contexto son sólo aplicables las aritméticas .

Y, para hallar las potencias de la variable X, habría que emplear cualquiera de las expresiones siguientes:

$$\begin{aligned} X^{**2} \\ X^{**(1/2)} \\ \ln(X) \\ -X^{**(-1)} \end{aligned} \quad (49)$$

La primera tiene el efecto de convertir en simétrica distribuciones asimétricas a la izquierda, pues $q > 1$, mientras las tres últimas, realizan el efecto contrario, esto es, reducir la asimetría de distribuciones con esta característica a la derecha.

Para las transformaciones lógicas, deberá utilizarse la instrucción *Transformar/Recodificar*, que, a su vez, está dividida en dos, según se desee realizar la operación sobre la misma o sobre distinta variable, siendo ésta la más razonable por las razones ya anteriormente expuestas en el caso de las transformaciones aritméticas.

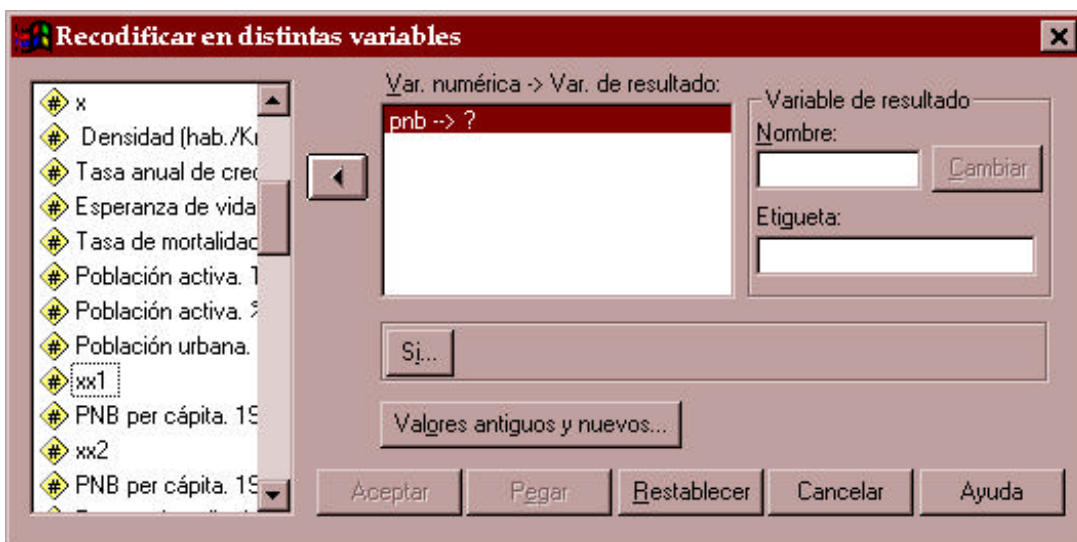


Figura 37.- Cuadro de diálogo para la recodificación de una variable.

La recodificación se realiza a través de dos pantallas de introducción de parámetros: En la primera (figura 37) se extrae una o varias variables ya existentes y listadas en la columna de la izquierda y debe proporcionarse un nombre nuevo a la variable que se creará tras la ejecución de la instrucción. La segunda pantalla (figura 38) es donde se especifica qué valores de la variable antigua se van a convertir en qué valores de la nueva variable. En el ejemplo de la figura se ha utilizado la variable PNB (Producto Nacional Bruto), cuyos valores oscilan entre un mínimo de 800 \$ y un máximo de 2.252.343 \$. Se ha deseado formar cuatro intervalos de desigual amplitud que reflejara una clasificación de los países en bajo, medio bajo, media alto y alto PNB. En la primera categoría se incluyen los países con menos de 5.000 \$ de PNB, en la segunda los que lo tienen entre 5.000 \$ y 50.000 \$, en la tercera los que se encuentran entre 50.000 \$ y 500.000 \$ y finalmente, la cuarta categoría incluye aquellos con un PNB superior a los 500.000 \$.

Figura 38.- Cuadro de diálogo para la recodificación de valores.

El cuadro de diálogo mostrado en la figura 38 está dividido en dos mitades verticales. En la de la izquierda aparecen siete botones correspondientes a siete modos de convertir los valores de la variable original. En primer lugar, se puede cambiar valor por valor, en cuyo caso habría que marcar el primer botón (*Valor*) e indicar en su correspondiente casilla el dato que se desea cambiar. También se puede transformar un conjunto de valores al mismo tiempo. Para ello, se utiliza el cuarto botón (*Rango*) que se compone de dos rectángulos, uno para indicar el valor mínimo del rango y el otro para exponer el valor máximo. El quinto y el sexto botón realizan sustancialmente la misma función que el anterior; pero los límites del rango vienen definidos en el primer caso por el valor mínimo de la variable y en el segundo por el valor máximo. Quedan tres botones a la izquierda del cuadro de diálogo comentado. El segundo y el tercero están para cambiar los valores perdidos de la variable original; el séptimo y último para especificar que se desean cambiar el resto de valores.

Cada una de estas posibilidades sirve para añadir una línea de cambio, que consiste en la transformación de un valor o rango de una variable en un determinado valor de la otra variable. Al menos deben realizarse tantas líneas como valores se desea que tenga la nueva variable. Como, en el caso del PNB, se deseaban cuatro valores, cuatro son las líneas necesarias. La primera convierte en 1 el rango desde el valor mínimo hasta el 5000; la segunda desde el valor 5000 hasta el 50.000; la tercera desde el 50.000 hasta el 500.000 y, finalmente, la cuarta transformación (aún pendiente en la figura, pues para que aparezca en el recuadro blanco de la derecha hay que pulsar el rectángulo *Añadir*) otorga a la nueva variable el valor 4 en el caso de que la original tenga un valor por encima de 500.000\$.

Por último, puede ponerse un ejemplo de variable nominal convertida a variable ficticia. En la ya comentada variable *estado civil*, si se desea convertir el valor representado con un "4" (divorciado) en la variable original, se podría realizar el cambio mediante dos líneas de

recodificación. En la primera, el "4" se convierte a un "1" y en la segunda *Todos los demás valores* se transforman a 0. (Figura 39).

Recodificar en distintas variables: Valores antiguos y nuevos

Valor antiguo

Valor:

Perdido por el sistema

Perdido por el sistema o usuario

Rango:

hasta

Rango:

Del menor hasta

Rango:

hasta el mayor

Todos los demás valores

Valor nuevo

Valor: Perdido por el sistema

Cogiar valores antiguos

Antiguo --> Nuevo:

Añadir

Cambiar

Borrar

Las variables de resultado son cadenas Ancho:

Convertir cadenas numéricas en números (5->5)

Continuar Cancelar Ayuda

Figura 39. Cuadro de diálogo para recodificación de variables ficticias.