

9

EL ANÁLISIS DE LOS DATOS

Una vez que ha concluido la recogida de información, comienza la fase de *análisis e interpretación de los datos*. Ésta es una fase compleja e igualmente crucial en el proceso de investigación. Si bien, en los últimos años se ha visto beneficiada por los grandes avances producidos en el campo de la informática. En concreto, la expansión de paquetes estadísticos informáticos (adaptados a ordenadores personales), que cubren el amplio espectro analítico existente (tanto *cuantitativo* como *cualitativo*). La implantación de estos programas ha adquirido tal relieve que, en la actualidad, sería inconcebible un *análisis de datos* ajeno al procesamiento informático.

En las páginas siguientes se ofrece una panorámica de las *técnicas cuantitativas de análisis de datos*. La notable pluralidad de éstas, unido a condicionantes de espacio, limitan su exposición a una mera descripción de rasgos distintivos de cada una de ellas. Para un mayor conocimiento y profundización, se remite a la consulta de bibliografía especializada en *análisis de datos* (véase la selección de textos que figuran al final del capítulo, y en la bibliografía general al término del libro).

9.1. Los preliminares del análisis de datos

“El paso más importante antes de comenzar a diseñar un proceso de entrada de datos es determinar qué programas se utilizarán para analizar los datos y convenciones concretas respecto a los formatos de ficheros y de *missing data*, que pueden manejarse para esos programas” (Fowler, 1988: 128). De ahí que la exposición de las *técnicas estadísticas de análisis* comience por sus *preliminares* esenciales: la creación de los *ficheros de datos*, junto a la depuración de la información (con especial referencia a la “no-respuesta”).

9.1.1. La creación de los ficheros de datos

Cada programa informático comprende instrucciones concretas para la creación de *ficheros de datos* propios y *ficheros de definición de los datos*. No obstante, la mayoría de estos programas permiten el acceso a ficheros elaborados por otros programas.

Primero, se confecciona un *fichero de datos* para la introducción de la información obtenida durante la investigación. Los *ficheros de datos* presentan formas diversas, dependiendo del programa informático que se maneje. La más común es el fichero de datos en formato *ASCII* (*American Standard Code for Information Interchange*). Este se compone de filas y columnas de números. Cada *fila* representa los datos recabados de cada sujeto o caso observado. En las *columnas*, en cambio, figuran los valores de las variables correspondientes a cada caso (véase el Cuadro 9.1). Este tipo de formato se precisa para ejecutar la mayoría de los *paquetes estadísticos genéricos* (como el SPSS, el BMDP o el SAS).

CUADRO 9.1. Extracto de un fichero de datos.

```
001211611450021402032540210011011934212416421102111005312123211200502013
002312911340010292016332221012101321122216221210211120143223612520202041
003811612430024332016341110611114300311111120000221130271124413211311122
004911291150021402032530200143221400222122111003132122232040207110002110
005211911450021212011141110923322113002221121121101001104422342614001103
006311811460731242014542210923411230012111211202002202432232211300310012
007611711440021212017541111042311230012211222200001002432236211040053212
```

Los *paquetes estadísticos genéricos* también pueden, sin embargo, leer *ficheros de datos* creados por *bases de datos* (como el DBASE) u *hojas de cálculo* (como LOTUS). Estos ficheros incluyen nombres para las variables, además de información relativa a la amplitud de la variable y la *codificación* de los valores de “no-respuesta” (*missing data*).

Cuando los datos se han registrado en *formato ASCII*, hay que elaborar también un *fichero de definición de datos*. Este comprende instrucciones precisas para la lectura de los datos que componen el *fichero de datos*. Estas instrucciones incluyen información primordial sobre las variables analizadas en la investigación: *columnas* donde se hallan ubicadas en el *fichero de datos*; las *etiquetas* dadas a las variables y a sus atributos; así como la especificación de cómo se *codifican* los valores “*missing*” (o sin respuesta), para su tratamiento en el análisis de los datos.

En el Cuadro 9.2 figura un extracto de un *fichero de definición de datos* en SPSS. Comienza con el nombre del fichero donde se han grabado los datos. Prosigue con la relación de variables y el número de las columnas donde éstas se encuentran registradas. Después, se asignan *etiquetas* a las variables que lo precisen, al igual que a sus valores. En el SPSS, como en los demás programas informáticos, se dan al usuario instrucciones

para la correcta elaboración de estos ficheros. También se expone cómo transportar y convertir ficheros de otros programas, su lectura y demás especificaciones necesarias para el análisis de los datos.

CUADRO 9.2. Extracto de un fichero de definición de datos en SPSS.

```
Data list file='menores.dat'.
/id 1-7 anio 8 sexo 9 edad 10 nacional 11 etnia 12 anorma 13 alcohol 14
heroína 15 cannabis 16 estudio1 to estudio2 17-20 desescol 21 profepa 22
situacpa 23 apreviv 24 ambiente 25 fuga 26 compania 27.
Variable labels desescol 'desescolaridad'
/profepa 'profesion del padre'
/situacpa 'situación empleo padre'
/apreviv 'apreciación vivienda'.
Value labels sexo 1 'varon' 2 'mujer'
/nacional 1 'espanol' 2 'extranjero'
/estudio1 to estudio2 01 'primero EGB' 02 'segundo EGB' 03 'tercero EGB'
04 'cuarto EGB' 05 'quinto EGB' 06 'sexto EGB' 07 'septimo EGB'
08 'octavo EGB' 09 'FP' 10 'BUP' 11 'compensatoria' 12 'otros'
13 'analfabeto'.
Missing values anio to compania (0).
Frecuencias anio to compania /statistics=all /hbar.
Save outfile='menores.sys'.
```

Aparte de estos ficheros, pueden crearse *subficheros* específicos con objeto de facilitar los análisis. Especialmente, cuando el tamaño de la muestra es elevado y el *fichero de datos* originario adquiere un gran volumen.

Estos *subficheros* incluyen una *muestra de variables*, restringida a aquellas que sean de interés para la ejecución de análisis concretos. También, pueden representar una submuestra aleatoria de los casos observados en la investigación. En este último caso, la finalidad no sería tanto el ahorro de tiempo en el análisis de la información, sino la *validación* de los resultados estadísticos obtenidos de la otra submuestra de la muestra global. Así, por *ejemplo*, para la *validación de análisis multivariantes* (como el de *regresión*), se recomienda la división de la *muestra* en dos *submuestras*: una, para la consecución del modelo; y la otra, para su *validación*.

9.1.2. La depuración de la información

A la creación de ficheros sigue la *depuración de los datos*, como antesala del análisis. El investigador ha de identificar posibles *errores* cometidos en la *grabación de los datos*. A tal fin resulta conveniente solicitar al programa la relación de *frecuencias* de todas las *variables* introducidas en el *fichero de datos* ("*frequencies*" en SPSS, o "*univariate*"

en SAS, por ejemplo). Esta relación incluye un listado con todos los valores de cada variable, la frecuencia (absoluta y relativa) de cada uno de ellos, los casos sin respuesta (*missing data*), estadísticos univariados y las representaciones gráficas que se soliciten.

De esta relación, se observará si existen anomalías en los valores de las variables codificadas. Más concretamente, si alguna de ellas incluye valores ajenos al recorrido o rango definido de la variable.

Por ejemplo, si en la variable sexo, que se ha codificado con sólo dos opciones de respuesta (1 'varon' 2 'mujer'), aparecen casos con valores superiores a dos (3, 4, 5 u otro), éstos corresponderían a sujetos que han sido erróneamente codificados o grabados en el ordenador. Por lo que, habría que proceder a su comprobación y corrección posterior.

Algunos programas informáticos (como el SAS o el SPSS) proporcionan, además, especificaciones dirigidas a la depuración de los datos (tanto durante su introducción, como una vez concluida ésta).

Si se observan incongruencias en los valores de las variables (como la anteriormente ejemplificada), ha de procederse a su localización y corrección. Los errores pueden deberse a fallos en la introducción de los datos en el ordenador, pero no siempre. También pueden ser ocasionados por deficiencias en la recogida de información.

En el primer caso, la corrección resulta más viable e inmediata: se revisan los cuestionarios (u otro instrumento de recogida de datos que se haya empleado), hasta localizar los casos en que se han grabado mal los valores de las variables; posteriormente, se introducirían los códigos correctos de las variables correspondientes.

Pero, si los errores se deben a un mal registro de la información en el instrumento de medición, las posibilidades de corrección se restringen. La dificultad de contactar de nuevo con la fuente de información (las unidades muestrales) lleva, no a la transformación de códigos numéricos, sino a la eliminación de aquellos casos con datos incorrectos o inconsistentes. Éstos se sumarían a aquellos que originariamente no proporcionaron información alguna. De esta forma, se incrementaría el volumen de los llamados "missing values" (o valores con los que se codifican las respuestas en blanco o incorrectas, y que se dan por perdidos).

Cuando unos casos específicos presentan muchos "missing values" en la mayoría de las variables, suele decidirse su exclusión del fichero de datos (salvo que al investigador le interese el análisis y descripción de estos casos). Igualmente, si de alguna variable se obtiene escasa información (teniendo un elevado porcentaje de valores missing), también suele optarse por su exclusión para el resto de los análisis.

En general, antes de proceder al análisis de los datos, el investigador evalúa los porcentajes de respuesta (para cada variable), y los "outliers" registrados en la matriz de datos.

Por "outliers" se entiende cualquier observación o caso que muestre inconsistencia con la serie global de datos. Su identificación requiere la realización de análisis univariados (para cada una de las variables), tanto numéricos como gráficos. Además, la mayoría de los programas informáticos incluyen instrucciones específicas para la detección de "outliers".

En cuanto a la "no-respuesta", su evaluación resulta igualmente exigida. Diversos autores, como Bourque y Clark (1994), recomiendan la comparación de las caracte-

rísticas demográficas de la *muestra* con las correspondientes a la *población* de la que ésta procede. Para ello se emplea el *Censo de Población*, u otra fuente de datos estadísticos o estudio, que describa al conjunto de la *población*.

Si de este análisis se dedujese la no *representatividad* de la *muestra*, el investigador deberá establecer la magnitud de las diferencias entre la *población* y la *muestra*. Esto es importante para la delimitación de las posibilidades de *inferencia* de las *estimaciones muestrales*. En palabras de Arber (1993: 71):

“La capacidad para realizar inferencias de una muestra a una población se basa en el supuesto de que la muestra lograda no esté sesgada por la *no-respuesta*. En la medida en que aquellos que no responden difieran de forma significativa de aquellos que sí responden, el investigador tiene una muestra sesgada.”

9.2. El análisis estadístico univariable

En el diseño de la investigación ya se prevén los análisis a realizar con la información reunida en el desarrollo de la investigación. Aunque ha de matizarse que el *proyecto de análisis* no es inmutable. Depende, en gran parte, de la *cantidad* y *calidad* de los datos que se recaben. De ahí la importancia que adquieren, en cualquier indagación, los *análisis exploratorios*, como paso exigido y previo a la decisión de qué técnica analítica (*bivariable* y *multivariable*) se va a aplicar.

9.2.1. La distribución de frecuencias

En la *exploración de los datos*, primero se procede a un análisis exhaustivo de cada *variable* incluida en la *matriz de datos* (*análisis univariable*). Para cada una de las *variables* se calcula su *distribución* o *tabla de frecuencias*. La *tabla de frecuencias* –como puede verse en el Cuadro 9.3– incluye los distintos *valores* que presenta la *variable* (distribuidos en *clases* o *categorías*), acompañados por su *frecuencia* (es decir, el número de veces en que aparecen).

En la primera columna (encabezada por el rótulo “*value label*”) figuran los distintos atributos que componen la *variable*. La siguiente columna (“*value*”) muestra el *valor* dado a cada *atributo*. En la tercera columna (la denominada “*frequency*”) se hallan las *frecuencias absolutas*; o sea, el número de casos (de la *muestra*) que comparten cada uno de los *valores* de la *variable*.

Para conocer la importancia de cada *valor*, y a efectos comparativos, se obtienen las *frecuencias relativas* o *porcentuales*, que representan cada *valor* en el conjunto de la *muestra* (columnas 4.^a y 5.^a). Primero, se calculan los *porcentajes* para toda la *muestra*; segundo, exclusivamente para aquellos casos que han proporcionado información al respecto (columna de “*valid percent*” o porcentaje válido). En esta columna no se consideran, por tanto, los “*missing values*”.

Este desglose de casos, en función de si aportan o no información, permite conocer la proporción de “*no-respuesta*” de cada *variable*. Este conocimiento adquiere especial relevancia para posteriores análisis.

Por último, se calculan las *frecuencias relativas acumuladas* (“cum percent”), a partir de las *frecuencias* contenidas en la columna 5.^a. Estas *frecuencias acumuladas* denotan la proporción de casos (válidos) que se encuentran por debajo, o por encima, de un determinado *valor* de la *variable*.

Cuando la *variable* está medida a nivel de *intervalo*, se aconseja la previa *agrupación* de los *valores*. Ello facilitará su presentación en una *tabla de frecuencias* de menores dimensiones. A este respecto, algunos autores –como Bryman y Cramer (1995)– sugieren que el número de *categorías* diferenciadas esté comprendido entre 6 y 20. Argumentan que menos de 6 y más de 20 categorías pueden distorsionar la *forma* de la distribución de la *variable*.

Los *estadísticos univariantes* que figuran en el Cuadro 9.3 se comentan en la sección 9.2.3, dedicada a su exposición.

CUADRO 9.3. Ejemplo de una tabla de frecuencias para la variable “estudio” en SPSS.

Value label	Value	Frecuency	Percent	Valid Percent	Cum Percent
Primero EGB	1	8	1.1	1.1	1.1
Segundo EGB	2	8	1.1	1.1	2.2
Tercero EGB	3	31	4.3	4.3	6.6
Cuarto EGB	4	59	8.3	8.3	14.8
Quinto EGB	5	119	16.6	16.7	31.5
Sexto EGB	6	152	21.3	21.3	52.8
Septimo EGB	7	177	24.8	24.8	77.6
Octavo EGB	8	72	10.1	10.1	87.7
FP	9	27	3.8	3.8	91.5
BUP	10	7	1.0	1.0	92.4
Compensatoria	11	1	.1	.1	92.6
Analfabeto	13	53	7.4	7.4	100.0
.	0	1	.1	MISSING	
	TOTAL	715	100.0	100.0	
Primero EGB	8				
Segundo EGB	8				
Tercero EGB	31				
Cuarto EGB	59				
Quinto EGB	119				
Sexto EGB	152				
Septimo EGB	177				
Octavo EGB	72				
FP	27				
BUP	7				
Compensatoria 1					
Analfabeto	53				
Mean	6.566	Std Err	.091	Median	6.000
Mode	7.000	Std Dev	2.420	Variance	5.856
Kurtosis	1.644	S E Kurt	.183	Skewness	1.000
S E Skew	.091	Range	12.000	Minimum	1.000
Maximum	13.000	Sum	4688.000		
Valid Cases	714	Missing Cases	1		

9.2.2. Representaciones gráficas

La información contenida en una *tabla de frecuencias* también puede representarse de forma gráfica. Ello ayuda a la visualización global de la concentración, o dispersión, de los datos en la variable considerada.

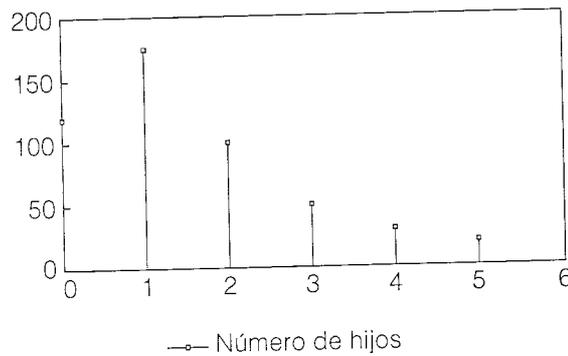
Dos de los *gráficos* habituales en la representación de *frecuencias* son el *diagrama de barras* y el *histograma*. A ellos se suman otros también usuales en el *análisis exploratorio*, como el diagrama de “*tronco y hoja*” y la “*caja*”; o el *polígono*, las *ojivas*, y el *gráfico de sectores*, entre la amplia variedad gráfica existente.

• *Diagrama de barras*

Consiste en una serie de “*barras*” (una para cada categoría de la variable), cuyas longitudes expresan las *frecuencias* de cada *atributo* de la *variable*.

EJEMPLO DE DIAGRAMA DE BARRAS

Número de hijos	Frecuencia absoluta
0	120
1	170
2	90
3	50
4	25
5	15
	470



• *Histograma*

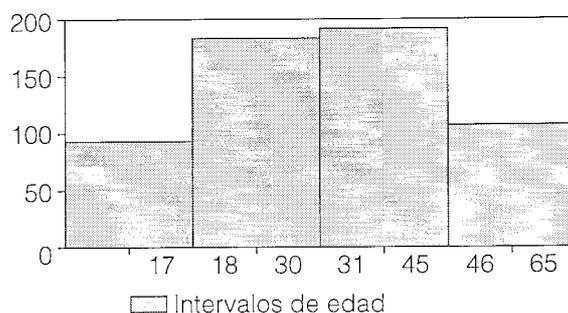
Un gráfico de contenido similar al anterior, aunque más adecuado a la representación de variables *métricas* (de *intervalo* o de *razón*).

En lugar de *barras*, el gráfico se compone de *rectángulos* unidos. Razón por la que se exige que la variable esté agrupada en *intervalos*. Estos forman las bases de cada uno de los *rectángulos*, que están delimitados por los *límites (reales)* de los respectivos *intervalos*. En cambio, la altura de los *rectángulos* es proporcional a la *frecuencia (absoluta o relativa)* de cada *intervalo*.

La suma total de las áreas de los rectángulos será igual a 1 (dado que la suma de todas las proporciones es la unidad).

EJEMPLO DE HISTOGRAMA

Edad	Frecuencia absoluta
Menos de 18	90
18 – 30	170
31 – 45	185
46 – 65	115
	560



• El “Tronco y las Hojas” (“Stem-and-Leaf”)

Constituye un gráfico parecido al *histograma*, pero integrado por los *dígitos* de los *valores* de las *variables*. Los *dígitos* se dividen entre dos. Los situados a la izquierda del punto (el *tronco*) figuran ordenados verticalmente, en orden creciente (de arriba a abajo). Por el contrario, los *dígitos* a la derecha del punto (las *hojas*), se disponen horizontalmente, aunque también en sentido creciente (de menor a mayor).

El *dígito* a la izquierda (la *columna*) que comprenda más *valores* a la derecha será aquél en el que se agrupan un mayor número de casos en la distribución. Por esta y otras razones, este gráfico suele tomarse como referente de las medidas de *tendencia central* de una distribución de frecuencias (a las que se hará referencia en el apartado 9.2.3).

EJEMPLO DEL GRÁFICO “EL TRONCO Y LAS HOJAS”

En el siguiente gráfico, puede observarse que la *fila* tercera representa los valores de mayor frecuencia en la distribución. En concreto, los *valores* 36, 36, 37, 37, 37, 38, 38, 39, 39, 39.

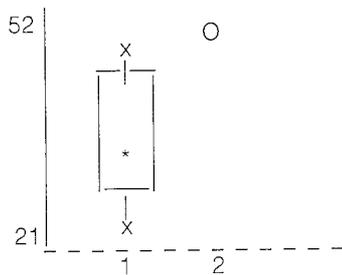
2	•	1 2 3 4
3	•	0 1 2 2 3 3 5
3	•	6 6 7 7 7 8 8 9 9 9
4	•	1 1 1 5 6 7 7
5	•	2 2

• La “Caja” (“Box-plot o Box-and-Whisker”)

Ofrece una visión global de la distribución, más sintética que la anterior. En ella la *variable* se representa de forma que el 50% de los casos queden comprendidos en el interior de la “caja”. En los extremos se sitúan, respectivamente, el 25% superior e inferior.

Con un asterisco se señala la *mediana*; con una “X” los valores máximos y mínimos; y con una “O”, los “*outliers*”. De esta manera se proporciona (gráficamente) información referente a la *mediana*, el *primer cuartil* (el 25% de los casos iniciales), el *tercer cuartil* (el 25% finales), y el *recorrido intercuartílico* (el 50% de los casos centrales) de la distribución de frecuencias. Ello exige que el nivel de medición mínimo de la variable sea el *ordinal*. En caso contrario, no podría estimarse el valor de la *mediana*, ni de ningún estadístico que precise de la ordenación de los *valores* de la variable, en un sentido creciente o decreciente (los *cuantiles*).

EJEMPLO DEL GRÁFICO LA “CAJA”



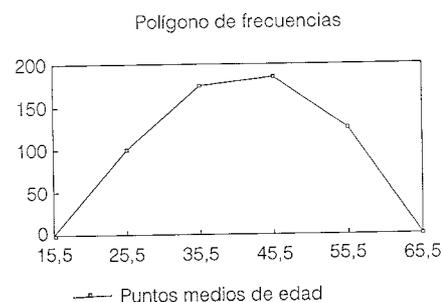
• Polígono

Un gráfico lineal que se traza sobre los *puntos medios* de cada *intervalo*, a una altura proporcional a su frecuencia (*absoluta* o *relativa*). Por lo que, de su visualización también se extraerán los *valores* con mayor (y menor) *frecuencia* en una distribución.

Los *puntos medios* se calculan sumando los *límites* de los *intervalos*, y dividiéndolo entre 2. De esta forma se obtiene el *valor* que representará al *intervalo* a todos los efectos. Por ejemplo, $(21 + 30)/2 = 25,5$.

EJEMPLO DE POLÍGONO DE FRECUENCIAS

Edad	Frecuencia absoluta	Punto medio
21 – 30	90	25,5
31 – 40	170	35,5
41 – 50	185	45,5
51 – 60	115	55,5
	560	



• *Ojivas*

Polígonos de frecuencias acumuladas que muestran la frecuencia de casos por encima, o por debajo, de un determinado valor de la distribución.

La *ojiva* será “menor que”, si se consideran los casos que hay por debajo de un valor. Por el contrario, será “mayor que”, cuando se representan los casos que comparten un valor superior de la distribución.

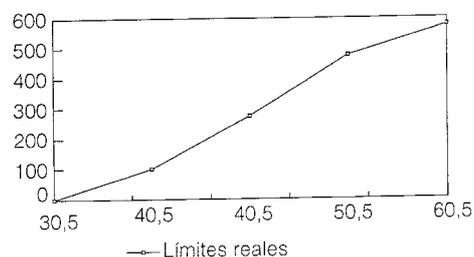
En el eje horizontal (de *abscisa*) se sitúan los *límites reales* de los intervalos de la variable, mientras que en el eje vertical (de *ordenada*) se disponen las *frecuencias acumuladas* (*absolutas* o *relativas*) de cada intervalo.

Para el cálculo de las *frecuencias acumuladas* se tiene en cuenta si interesa conocer el número (o la proporción) de casos que hay por debajo (*ojiva “menor que”*), o por encima (*ojiva “mayor que”*) de un valor específico de la distribución.

EJEMPLO DE OJIVAS “MENOR QUE” Y “MAYOR QUE”

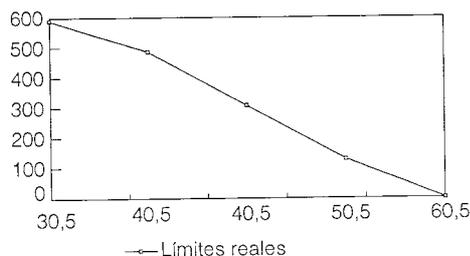
Ojiva “menor que”

Edad	Frecuencia absoluta	Frecuencia acumulada
21 – 30	90	90
31 – 40	170	260
41 – 50	185	445
51 – 60	115	560
	560	



Ojiva "mayor que"

Edad	Frecuencia absoluta	Frecuencia acumulada "más de"
21 - 30	90	560
31 - 40	170	470
41 - 50	185	300
51 - 60	115	115
	560	



• Gráfico de sectores (o en forma de "tarta")

A diferencia de los gráficos anteriores, éste se representa mediante un círculo, dividido en "sectores", cuyos ángulos indican el porcentaje de casos que comparten cada atributo de la variable.

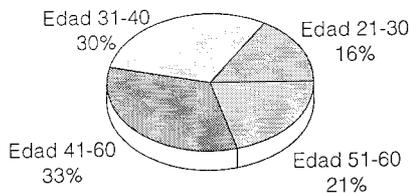
Para su obtención manual, se multiplican las *frecuencias relativas* (porcentual) de los *intervalos* (o valores) de la variable por 3,6 (que son los *grados* que corresponden a cada unidad porcentual: 360/100). De esta forma se transforman los *porcentajes* en *grados*, lo que posibilita su representación en una circunferencia.

No obstante, el empleo de programas gráficos (como el Harvard Graphics) hace innecesaria la realización de estos cálculos. Basta con conocer la *frecuencia* de cada *valor* de la distribución.

Los "sectores" del gráfico suelen representarse con colores o trazas diferentes, que proporcionen un mayor contraste entre ellos.

EJEMPLO DE GRÁFICO DE SECTORES

Edad	Frecuencia absoluta	Frecuencia relativa (%)	Grados
21 - 30	90	16	58
31 - 40	170	30	108
41 - 50	185	33	119
51 - 60	115	21	75
	560		360



Para un conocimiento más amplio de éstas y demás representaciones gráficas comunes en la investigación social, se remite a textos específicos sobre *Gráficos* (como el de Alaminos (1993), y otros citados en la bibliografía).

9.2.3. Estadísticos univariantes

Aparte de los gráficos, en la *exploración univariante* también se emplean estadísticos para medir, de manera precisa, la distribución de los valores de una variable (véase Cuadro 9.3). Su uso dependerá, sobre todo, del nivel de *medición* de la *variable*.

Los *estadísticos univariantes* se agrupan en tres grandes apartados:

- a) Medidas de tendencia central.
- b) Medidas de dispersión.
- c) Medidas de forma de la distribución.

• *Medidas de tendencia central*

Describen cómo se agrupan los *atributos* de una *variable* alrededor de un “valor típico” de la distribución. Por lo que, proporcionan una síntesis de la información contenida en la distribución.

Las medidas de *tendencia central* más empleadas en la investigación social son las siguientes: la *media*, la *mediana* y la *moda*.

- 1) La *media* es la medida más representativa, siempre y cuando la variable sea cuantitativa (de *intervalo* o de *razón*). Su cálculo precisa de la participación de todos los valores de la distribución. Cada uno de ellos se multiplica por sus respectivas *frecuencias absolutas*. Después, se suman todos los productos, y el resultado se divide por el número total de casos. De esta forma se obtiene el *promedio* de los valores de la distribución, que es como se define la *media aritmética*. El *inconveniente* fundamental de este estadístico es que se ve afectado por *valores* muy extremos en una distribución. Cuando la variable incluye *valores* muy dispares (o extremos), la *media* pierde representatividad a favor de la *mediana*, como *medida de tendencia central*.
- 2) La *mediana* es el valor que divide a la distribución en dos partes iguales. Se sitúa en el medio de la distribución. Lo que permite el conocimiento de los *valores* de mayor representación en la *muestra*. Su cálculo precisa que la variable sea, al menos, ordinal.
- 3) La *moda* denota el valor de mayor frecuencia en una distribución; aquel que más casos comparten. La distribución puede ser *unimodal* (una sola *moda*), pero también *bimodal* (dos *modas*), o *multimodal* (más de dos *modas*). Esto dificulta su interpretación, en caso de coexistir varias *modas*. A este *inconveniente* primordial se suma otro importante: en su cálculo no intervienen todos los *valores* de la distribución. Si bien, su principal *ventaja* es su universalidad. Pue-

de estimarse para cualquier tipo de variable, ya que el *nivel de medición* mínimo exigido es el *nominal*.

Además de estos estadísticos, existen los *cuantiles* como *medidas de tendencia no central*. Éstos representan *valores* que dividen a la distribución en partes iguales. Los *cuartiles* la dividen en cuatro partes iguales (cada una de ellas incluye al 25% de los *valores*); los *deciles* en diez partes; y los *percentiles* en cien partes. Su cálculo se asemeja al de la *mediana*. De hecho, el *cuartil* dos, el *decil* cinco, y el *percentil* cincuenta, expresan el *valor mediano*.

• Medidas de dispersión

Al conocimiento de los *valores centrales* de la distribución le sigue la medición de su representatividad: la mayor o menor *variabilidad* existente en torno a la *media* o la *mediana* de la distribución.

Las medidas de *dispersión absolutas* más comunes son el *rango* (o recorrido), la *desviación típica*, y la *varianza*.

- 1) El *rango* o *recorrido* expresa el número de valores incluidos en la distribución. Estos se obtienen de la diferencia entre el valor superior y el inferior. Su comprensión es sencilla, aunque presenta, en su contra, un *inconveniente* importante: es susceptible a la distorsión proporcionada por valores muy extremos en una distribución. Ello se debe a la única consideración del primer y último valor de la distribución.

Para solventar esta deficiencia, suelen aplicarse otras medidas de *rango*, que emplean un mayor volumen de información. Si bien, éstas exigen que la variable sea, al menos, *ordinal*. Se trata de los *recorridos*: *intercuartílico* (la diferencia entre el tercer *cuartil* y el primero; por lo que incluye al 50% de los valores centrales de la distribución), *semiintercuartílico* (el 25% de los valores centrales), *interpercentílico* (el 80% de los valores centrales, al ser la diferencia entre el *percentil* 90 y el 10), y el *semiinterpercentílico* (el 40%).

- 2) La *desviación típica* es el promedio de la desviación de los casos con respecto a la *media*. Como indicador de heterogeneidad (o de dispersión de los valores de una distribución), su estimación se exige siempre que se calcule la *media*, porque ayudará a la interpretación de su representatividad en la distribución.

Al igual que la *media*, su valor viene expresado en la unidad de medición de la *variable*, y únicamente puede calcularse cuando la variable es *cuantitativa*.

- 3) La *varianza* constituye otra medida de heterogeneidad de una distribución. Se define como el cuadrado de la *desviación típica*. Su valor expresa el grado de heterogeneidad de una población respecto a la *variable* medida, siendo sus características similares a la *desviación*.

A partir de la *desviación*, puede calcularse una medida de *dispersión relativa* que ayudará en su interpretación. Se trata del *coeficiente de variabilidad de Pearson*. Una medida estandarizada que se obtiene del cociente entre el valor de la *desviación* y la *media aritmética*. Su valor se expresa en porcentajes, siendo de utilidad en la comparación de la homogeneidad de dos o más grupos respecto a una o más variables.

Por último, cuando se calcula la *mediana*, también suelen estirmarse estadísticos que midan su representatividad en la distribución. Concretamente, la *desviación media de la mediana* y el *coeficiente de variabilidad de la mediana de Pearson* (obtenido del cociente entre la *desviación media de la mediana* y la *mediana*). Su interpretación es análoga a los estadísticos anteriores.

Para un conocimiento más detallado, remito a la consulta de cualquier manual de estadística.

- *Medidas de la forma de la distribución*

Este término comprende la disposición espacial de los *valores* en una distribución: su *asimetría* y *curtosis*.

- 1) La *asimetría* constituye un indicador de la *agrupación* de las *frecuencias* en la *curva* de una distribución. Cuando su valor es cero, expresa que la *curva* es *simétrica*; es decir, que coinciden los valores de la *media*, la *moda* y la *mediana*. Si es mayor de cero, significa que la distribución es *asimétrica a la derecha* (o *de sesgo positivo*). Los casos se agrupan a la izquierda de la *curva*. Esto significa la mayor presencia en la distribución de *valores* inferiores a la *media*.

En cambio, si el valor del coeficiente de sesgo es inferior a cero, denota que la agrupación de los *valores* se produce a la derecha de la *curva*. Por lo que habría una mayor representación de los *valores* superiores a la *media*.

En general, un valor superior a 0.8 (positivo o negativo) indica que la *asimetría* de la variable es importante.

- 2) La *curtosis* hace referencia a la mayor o menor *concentración* de *valores* en torno a la *media* de la distribución. Si existe una elevada concentración, la distribución será *leptocúrtica*. Esto significa que la *media* es muy representativa, al haber muy poca dispersión respecto a ella.

La distribución será *platicúrtica*, en el caso contrario: cuando existe una elevada *dispersión* de *valores* respecto a la *media*. Lo que expresa su escasa representatividad.

Con estos estadísticos lo que se pretende es comprobar si existen valores muy extremos en una distribución. En caso de existir, puede decidirse su transformación, con la finalidad de alcanzar una distribución que se aproxime a la *normal* (*simétrica* y *mesocúrtica*).

Tras los primeros *análisis univariados*, el investigador adquiere una descripción inicial de cada una de las *variables* que componen la investigación. Este conocimiento le

puede llevar a introducir *modificaciones en la configuración de las variables*, de forma que se faciliten los análisis posteriores. Estas *modificaciones* consisten, esencialmente, en agrupaciones de *variables* o de sus *atributos*.

Cuando en una distribución se observa que una o varias *categorías* presentan una *frecuencia* muy baja, habría que proceder a su *agrupación* con otras *categorías* de similar significado (si se pretende, con ellas, efectuar *análisis bivariantes* o *multivariantes*). Si la variable es *nominal*, la agrupación se produce con *categorías* que presentan un nexo común. Pero, si la variable es *ordinal*, de *intervalo*, o de *razón*, han de agruparse sólo los *valores* que se hallen más próximos en la escala (por *ejemplo*, las categorías “extrema derecha” con “derecha”).

A la nueva *categoría* (formada de la agrupación de dos o más categorías originales) habría que darle una nueva denominación, que resuma la variedad de *atributos* que comprende. De este modo las posibilidades de análisis, de variables con *atributos* de escasa representación en la *muestra*, se amplían; aunque la agrupación lleve consigo pérdida de información.

Igualmente, puede *agruparse variables* similares, con la finalidad de componer una medida única que sintetice la información contenida en *variables* análogas.

9.3. El análisis bivariante

Después de la realización de los *análisis exploratorios*, procede la realización de *análisis bivariantes*, tanto con fines *descriptivos* (describir al conjunto de la población observada), como *explicativos* (analizar posibles relaciones causales entre dos variables: la independiente y la dependiente).

9.3.1. Las tablas de contingencia

En la investigación social, la práctica habitual es la confección de *tablas de contingencia*, formadas del cruce, al menos, de dos *variables*. Estas *tablas* generalmente se obtienen mediante los comandos CROSSTABS y TABLES, en la mayoría de los programas estadísticos.

En estos comandos se especifican las *variables* a cruzar. Si puede establecerse una *relación causal* entre ellas, el orden convencional de exponer las *variables* es: *dependiente* “by” *independiente*. De esta forma, la *variable dependiente* figurará en las *filas*, mientras que la *independiente* en las *columnas*.

Si se desean introducir *variables de control* (para la eliminación de *explicaciones alternativas*), estas terceras y, a veces, cuartas variables, se añaden a las anteriores, siendo igualmente precedidas por la preposición “by” (por *ejemplo*, CROSSTABS = EDAD BY HERMANOS; CROSSTABS = MEDIDA BY DELITO BY SEXO). No obstante, en el programa informático que se emplee, se especifican las instrucciones propias para la ejecución del comando.

Además de las *variables*, ha de delimitarse la información que se precise: *frecuencias absolutas*, *porcentajes* (horizontales, verticales, totales), y *estadísticos* que midan el grado y la significatividad de la relación entre las *variables*.

El Cuadro 9.4 muestra una *tabla de contingencia simple* obtenida mediante el comando CROSSTABS de SPSS. En él se solicitaron las *frecuencias absolutas* ("count"), y las *relativas*: porcentajes horizontales ("row") y verticales ("column"), para cada *casilla* de la *tabla*. A ello se sumaron los *estadísticos de contingencia*.

CUADRO 9.4. Una tabla de contingencia simple mediante SPSS.

Crosstabulation:		Delito By medida			Row total
Medida	Count Row Pct Col Pct	Amones- tación 1	Libertad vigilada 2	Interna- miento 3	
Delito	1	40 51.3 19.3	26 33.3 11.7	12 15.4 12.2	78 14.8
Hurto	2	12 19.7 5.8	36 59.0 16.1	13 21.3 13.3	61 11.6
Intimidación con armas	3	116 45.8 56.0	93 36.8 41.7	44 17.4 44.9	253 47.9
Robo sin intimidación	4	10 13.7 4.8	43 58.9 19.3	20 27.4 20.4	73 13.8
Insumisión paterna	5	29 46.0 14.0	25 39.7 11.2	9 14.3 9.2	63 11.9
Otras infracciones					
	Column total	207 39.2	223 42.2	98 18.6	528
	<u>Chi-Square</u>	<u>D.F.</u>	<u>Significance</u>	<u>Min E.F.</u>	<u>Cells with E.F.< 5</u>
	41.32788	8	.0000	11.322	None
	<u>Statistics</u>	<u>Symmetric</u>	<u>With DELITO Dependent</u>	<u>With MEDIDA Dependent</u>	
	Lambda	.07069	.00000	.13443	
	Uncertainty Coefficient	.03488	.03033	.04102	
	Somers'D	.04951	.05219	.04710	
	Eta		.07227	.17561	
	<u>Statistics</u>	<u>Value</u>	<u>Significance</u>		
	Cramer's V	.19783			
	Contingency Coefficient	.26943			
	Kendall's Tau B	.04958	.0952		
	Kendall's Tau C	.04959	.0952		
	Pearson's R	.04051	.1764		
	Gamma	.07346			
Number of Missing Observation = 0					

Cada *casilla* es el resultado del cruce de una *fila* con una *columna* (es decir, del cruce de un *atributo* de una *variable* con el *atributo* de la otra *variable*).

La lectura de las *tablas* con frecuencia se limita a comentarios porcentuales. Se contrastan los *porcentajes* de cada *casilla* para comprobar la existencia de variaciones entre los distintos *atributos* de las *variables*. A tal fin se calculan los *porcentajes* a partir de los *marginales* de la *tabla*.

Si se toma como base el total de *filas*, el *porcentaje* será *horizontal*, y las comparaciones porcentuales (entre los subgrupos) se efectuarán verticalmente. Por el contrario, cuando la base la constituye el total de *columnas*, se procede a la inversa: el *porcentaje* será *vertical*, y las comparaciones de porcentajes en sentido horizontal.

El investigador deberá escoger entre uno u otro tipo de *porcentaje* (horizontal o vertical), en conformidad con los *objetivos* del estudio y las *hipótesis* que compruebe. Si bien, existen mayores restricciones en *estudios explicativos* que en los *descriptivos*.

Si la finalidad de la investigación es la búsqueda de *relaciones causales*, los *porcentajes* se estiman sólo en el sentido de la variable *independiente*. Esta variable suele situarse en las *columnas*, y la *dependiente* en las *filas*; salvo que el elevado número de *atributos* de la variable *independiente* desaconseje su ubicación en las *columnas*. Esta disposición de las variables responde a la mayor facilidad de lectura (en la cultura occidental) en sentido horizontal frente al vertical. Los *porcentajes* serían, por tanto, *verticales* y las comparaciones horizontales.

Las diferencias porcentuales deben superar un determinado valor (al menos superior al 5%) para que puedan considerarse importantes. Depende del *error muestral* que derive de los tamaños de las *bases* sobre las que se calculan los *porcentajes*. Si estos *tamaños muestrales* son bajos, la diferencia porcentual ha de ser superior, si de ella quiere deducirse la existencia de *asociación* entre las *variables*.

En la exposición de la *tabla* (en el informe de la investigación) ha de indicarse, explícitamente, la dirección en la que se han calculado los *porcentajes*. Como sólo se aportan datos porcentuales, se recomienda poner entre paréntesis las *bases* de los *porcentajes*. Ello ayudará a la interpretación de las diferencias que en ellos se observen.

También se aconseja encabezar la *tabla* con un *título* que describa, sucintamente, el contenido de la *tabla*. En el *título* han de especificarse las *variables* comprendidas en la *tabla* y su relación.

A modo de ejemplo, la *tabla de contingencia* expuesta en el Cuadro 9.4 puede transformarse, en el *Informe*, como se expone en el Cuadro 9.5. Si bien, ha de matizarse que ésta constituye una de las posibles alternativas.

Pese a la existencia de convencionalismos en el formato de las tablas, el investigador es libre de diseñar el formato que más se ajuste a su estilo particular, y al contenido de la *tabla*.

La *tabla* expuesta en el Cuadro 9.5 constituye una *tabla simple* porque en ella figuran sólo dos *variables*. Pero también suelen componerse *tablas complejas*, a partir de

CUADRO 9.5. Menores clasificados por medida del tribunal, según el tipo de delito (porcentaje vertical).

Medida	Tipo de delito					Total
	Hurto	Robo con intimidación	Robo sin intimidación	Insumisión paterna	Otros delitos	
Amonestación	51	20	46	14	46	39
Libertad vigilada	33	59	37	59	40	42
Internamiento	16	21	17	27	14	19
Total	100 (78)	100 (61)	100 (253)	100 (73)	100 (63)	100 (528)

la conjunción de varias *variables* (independientes y/o dependientes). En estos casos, los análisis se centran en las variaciones en los *valores* extremos de las *variables*.

La *lectura porcentual*, aunque ilustrativa, resulta, no obstante, insuficiente. Precisa del complemento de *estadísticos* que gradúen la *asociación* entre las *variables* y su *significatividad*. Este complemento adquiere un mayor protagonismo cuando de las *tablas* quiera deducirse una *relación causal*.

Dos *variables* se hallan relacionadas si sus *atributos* varían conjuntamente. Para la graduación de esta relación, se acude a alguno de los *estadísticos de contingencia*, dependiendo del nivel de medición de la variable. Aquí sólo se enumeran. Consúltense en un manual de estadística su formulación.

- a) Si la variable es *nominal*, los estadísticos que miden el grado de *asociación* entre dos *variables* son: Phi cuadrado, "C" de Pearson, "V" de Cramer, "Q" de Yule, Lambda, Tau-Y de Goodman y Kruskal...; además del coeficiente "d" o de diferencia de proporciones.
- b) *Variables ordinales*: Rho de Spearman, Tau-A, Tau-B y Tau-C de Kendall, Gamma de Goodman y Kruskal, "D" de Sommer, entre otros.
- c) *Variables de intervalo*: a los estadísticos anteriores se suma el coeficiente de correlación producto-momento de Pearson.

Cada uno de estos estadísticos indican la fuerza y la dirección de la *asociación* entre dos variables. Su *signo* expresa la dirección de la *correlación* (positiva o negativa); mientras que el *valor* numérico (que oscila entre 0 y 1), la magnitud de la relación. De este modo:

- a) El valor ".00" denota *inexistencia de asociación*.
- b) "-1.00", *correlación perfecta negativa* (conforme aumenta el valor de la variable independiente, disminuye el valor correspondiente a la dependiente).

- c) “1.00”, *correlación perfecta positiva* (al incremento de la variable independiente le sigue el aumento, también, de la dependiente).

Una vez medida la relación entre las variables, se comprueba su *significatividad*. Los datos analizados siempre pertenecen a una *muestra*, de las múltiples posibles, que pueden extraerse de una misma *población*. Razón por la cual, se exige la comprobación de la *significatividad* de los estadísticos y sus posibilidades de *inferencia* a la *población*.

En las *tablas de contingencia*, se aplica el *test de la Chi-Cuadrado* (X^2). Este estadístico se obtiene de la comparación entre las *frecuencias observadas* (en la *muestra*) y aquellas que cabría esperar en caso de inexistencia de relación entre las variables. Su valor se compara con el *teórico* (aquel que figura en una tabla de la X^2), para unos grados de libertad determinados y un nivel de significación escogido por el investigador. El *nivel de significación* habitual es .05. Este supone una posibilidad de error en la estimación del 5%. Los *grados de libertad*, en cambio, vienen marcados por las dimensiones de la *tabla*: número de *filas* (i) y de *columnas* (j). Concretamente, $g. l. = (i-1)(j-1)$.

Cuando el valor de la X^2 empírico (el obtenido en la *muestra*) supera al teórico (el marcado en una tabla de la X^2), se deduce la *significatividad* estadística de la relación observada entre las *variables*. En caso contrario (X^2 empírico < X^2 teórico), se desestima la relación bivariable, por su no *significatividad*. La relación se consideraría, entonces, meramente casual, debida a *errores muestrales*.

En la salida de ordenador (véase Cuadro 9.4, por ejemplo) el valor de la X^2 aparece acompañado con su *significatividad* (“*significance*”). El valor que figura bajo este rótulo ha de ser inferior a .05 para que el valor de la X^2 sea significativo a un nivel de .05.

En cualquier manual de estadística en las ciencias sociales (como el de Blalock, 1978; García Ferrando, 1985; o Siegel, 1985) puede encontrarse una exposición detallada de estos y demás *estadísticos de contingencia*.

9.3.2. Otros análisis bivariantes

Aparte de las *tablas de contingencia*, existen otras *técnicas de análisis bivariantes*, como el de *regresión* y *varianza simple*. Ambas técnicas analíticas miden la relación de *dependencia* entre dos variables, si bien imponen mayores restricciones que las *tablas de contingencia*. Su cumplimiento exige que la *variable dependiente* sea *métrica* o cuantitativa. Ello determina su menor aplicabilidad en la investigación social, en la que predominan las *variables cualitativas* (*no métricas*).

• *Varianza simple*

Esta técnica analítica es muy aplicada en los *diseños experimentales*, en la comprobación de los efectos de los tratamientos experimentales. A ello contribuye la es-

pecificidad del análisis. Su finalidad es comprobar la existencia de diferencias grupales respecto a una única *variable dependiente (métrica)*. Para ello se manipula una *variable independiente*, en función de cuyos valores se forman distintos grupos de tratamiento. Constituidos los grupos, se comprueba la media de cada uno de ellos respecto a la *variable dependiente*. Si se observan diferencias entre las *medias grupales*, se procede después a la comparación de las *varianzas grupales*, y a la medición de su *significatividad*.

Interesa que la *varianza* entre los grupos supere a la *varianza intragrupal*. Ello expresaría una mayor heterogeneidad entre los grupos, frente a una escasa variabilidad dentro de ellos. Por lo que podría afirmarse la existencia de diferencias entre los grupos.

La *significatividad* de las diferencias grupales se comprueba mediante los estadísticos “t” (si únicamente se han formado dos grupos de tratamiento), y “F” (si son más de dos los grupos creados).

En ésta, como en cualquier prueba de *significatividad*, se comparan los *valores empíricos* (“t” y “F”) con los *teóricos* (mostrados en las tablas de la “t” de Student y de la “F” de Fisher, correspondientes). El proceso es similar al descrito en el *test de la Chi-Cuadrado*. Se fija el *nivel de significatividad*, en función de la precisión que el investigador desee para su estimación (.05, generalmente); y los *grados de libertad* (ahora determinados por el *tamaño muestral* y el número de *variables independientes* consideradas). Siempre que el *valor empírico* supere al *teórico*, las diferencias grupales observadas en la *muestra* adquirirán *significatividad estadística*. Podrán, por tanto, hacerse extensibles al *universo* del que se extrajo la *muestra* (en los niveles de probabilidad fijados).

• Regresión simple

Constituye otra técnica de *dependencia* en la que se analiza la relación entre una única *variable independiente (métrica o no métrica)* y una *dependiente (métrica)*. Pero, a diferencia de la técnica analítica anterior, la finalidad del análisis es la *predicción* del valor de la *variable dependiente* a partir del conocimiento de la *independiente*. Se cuantifica la relación existente entre ambas variables; y, se establece el grado de confianza o *significatividad* de la estimación efectuada.

La *correlación* entre las dos variables (*dependiente e independiente*) se mide mediante el *coeficiente R de Pearson*. Éste expresa el grado de *covariación* entre las variables, según se aproxime a “0” (inexistencia de asociación) o a “1” (asociación perfecta). También informa de la dirección de la asociación: creciente (si el signo es positivo) o decreciente (si es negativo).

En el *análisis de regresión*, la idea que subyace es la consecución de una *recta de regresión* que presente el mejor “ajuste” de los casos respecto a las variables analizadas. Esta *recta* tiene su expresión matemática en la siguiente *ecuación de regresión*:

$$y = a + bx + e$$

donde: “y” denota el valor de la variable *dependiente*.

“a” es el *intercepto* o punto de la *recta* que corta al eje de las Y.

“b” es la *pendiente* de la *recta* (también referido como el *coeficiente de regresión*). Su *valor* expresa la cantidad de variación de la variable *dependiente* por cada unidad de variación de la *independiente*. Su *signo* denota si se produce aumento (pendiente creciente; signo positivo) o disminución (pendiente decreciente; signo negativo).

“e” representa el *error de la estimación*: la inadecuación de la *ecuación de regresión* en la predicción del valor de la variable *dependiente*.

Esta *ecuación* permite la predicción del valor de la variable *dependiente* a partir de valores conocidos de la *independiente*. Los *coeficientes* se obtienen, generalmente, siguiendo el *criterio de mínimos cuadrados* (hacer mínima la distancia que separa los *puntos* –obtenidos de la confluencia de ambas variables en cada uno de los casos– y la *recta de regresión*).

La *significatividad* de los *coeficientes* se comprueba mediante el estadístico “t”, con n-1 *grados de libertad* (siendo “n” el número de observaciones). En cambio, la *significatividad* de la *correlación* se comprueba mediante el estadístico “F”. Como en cualquier prueba de *significatividad*, los valores de “t” y de “F” *empíricos* han de superar los *teóricos* (determinados en las tablas respectivas) para que el modelo de *regresión* sea significativo estadísticamente.

9.4. El análisis multivariable

Los análisis *univariantes* y *bivariantes* con frecuencia se muestran insuficientes para cubrir los objetivos de la investigación. El proporcionar una visión conjunta e integrada, que describa y/o explique la realidad que se analiza, demanda la realización de *análisis multivariantes* (de más de dos variables al mismo tiempo). De otra forma no podrían medirse las influencias e interrelaciones existentes entre grupos de variables. Como hace tiempo reconociera García Ferrando (1979: 198):

“Las distribuciones bivariantes en sociología aparecen demasiado simplistas para lograr adecuadas explicaciones científicas.”

La peculiaridad del *análisis multivariable* reside en operar con un número elevado de variables, y de manera simultánea, basándose en el cálculo matricial. Kendall (1975) lo define como el conjunto de técnicas estadísticas que permite el análisis simultáneo de más de dos variables en una muestra de observaciones. A esta definición, Dillon y Goldstein (1984) añaden la posibilidad de análisis sincrónicos de mediciones en más de una *muestra*.

Para el *análisis multivariable* existe un amplio abanico de *técnicas*. En conformidad con un extenso grupo de autores (véase Kendall, 1975; Dillon y Goldstein, 1984; o Hair *et al.* 1992, por ejemplo), estas técnicas pueden agruparse en dos grandes categorías (*técnicas de dependencia* y *técnicas de interdependencia*), en función de si se diferencia, o no, entre variables *dependientes e independientes*.

En la elección de la *técnica* concreta a aplicar intervienen, básicamente, el objetivo de la investigación, y las características de las variables que se analicen (su número y nivel de medición).

A continuación se ofrece una breve descripción de cada una de las *técnicas multivariadas* comúnmente referidas. Para un mayor conocimiento de cada una de ellas remito a la bibliografía especializada. Lo que sigue es una mera visión panorámica de las *técnicas multivariadas*.

9.4.1. Técnicas multivariadas de dependencia

Un conjunto de técnicas analíticas unidas por un mismo propósito: medir la existencia de *relaciones causales* entre un conjunto de variables, el grado y *significatividad* de la misma. Sin embargo, difieren en el número de variables *dependientes* que incluyen, y en el nivel de *medición* exigido (*métrico o no métrico*). En la Figura 9.1 se esquematizan las principales alternativas, comúnmente señaladas, en el análisis de la *dependencia*.

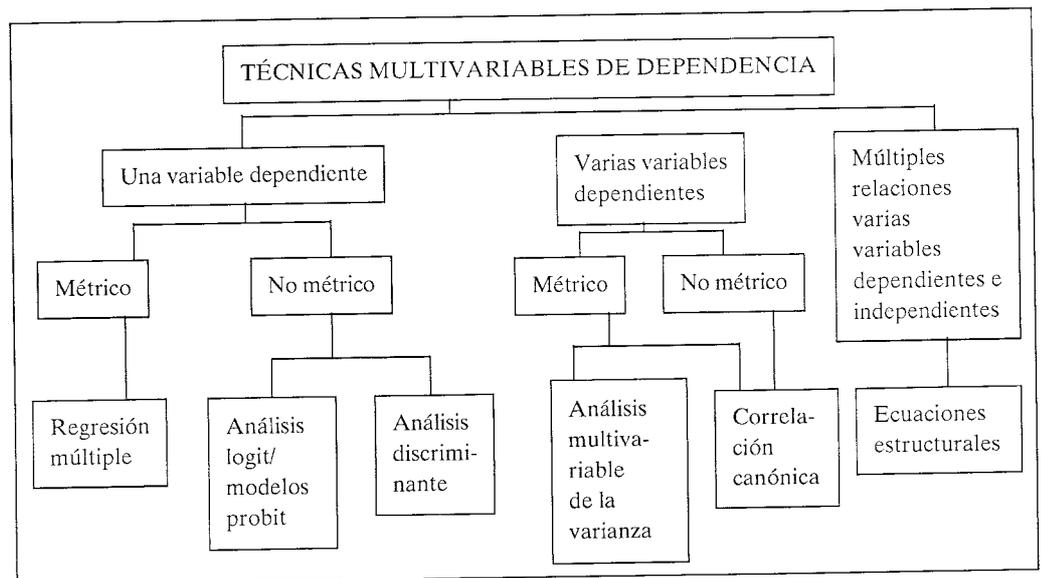


Figura 9.1. Técnicas de análisis multivariable de dependencia.

- *Regresión múltiple*

Una técnica de *dependencia* muy versátil, y utilizada, cuando se busca la *predicción* del valor de una variable *dependiente* (o *criterio*), a partir de valores conocidos en una serie de variables *independientes* (o *predictoras*). Para su realización se exige el cumplimiento de una serie de *supuestos* básicos:

- Linealidad*: la interdependencia entre las variables ha de responder a un modelo lineal. Esto quiere decir, que a cada cambio en una unidad en una variable corresponda un cambio, también en una unidad, en la otra variable.
- Normalidad*: la correspondencia de la distribución de los datos (para la variable *dependiente* y las *independientes*) con la *curva normal*. Esto permite la utilización de los estadísticos “*F*” y “*t*” para la comprobación de la *significatividad*.
- Homocedasticidad* o igualdad de las *varianzas* de los términos de error en la serie de variables *independientes*.

Para que pueda medirse la relación de las variables *independientes* con la *dependiente* se precisa que la *varianza* de los valores de la variable *dependiente* sea igual en cada valor de las variables *predictoras*.

- Aditividad*: los efectos de las variables *independientes* han de poderse sumar entre sí, para poder predecir la variable *dependiente*.
- Ausencia de *colinealidad* (de correlación) entre las variables *independientes*, con objeto de que puedan medirse sus efectos concretos en la variable *dependiente*.
- Ha de haber un número elevado de observaciones. Al menos, debería haber 20 veces más casos que variables *independientes*. Máxime si el modelo de regresión se obtiene “paso a paso”: incorporando una a una las variables *independientes* hasta que no exista ninguna más con poder predictivo significativo.

Cuando no se obtiene esta proporción, habría que optar por alguno de los siguientes remedios: eliminar alguna variable *independiente*, o agrupar varias variables creando una nueva variable (que sea una combinación de dos o más variables *independientes*).

Como en *regresión simple*, en *regresión múltiple* se obtiene una *ecuación de regresión*, con tantos *coeficientes* como variables *independientes*. A partir de ella, podrá predecirse el valor medio de la variable *dependiente*. Asimismo, se mide el grado de *correlación* existente entre las variables (mediante el estadístico *R múltiple*) y su *significatividad* (gracias al estadístico “*F*”).

El valor del R^2 *múltiple* (el *coeficiente de determinación*) expresará el porcentaje de *varianza* de la variable *dependiente* que es explicado por las *independientes*. Cuanto más elevado sea su valor, mejor para la predicción de la variable *dependiente*.

- *Análisis logit*

Los análisis *logit* y *probit* se definen como modelos analíticos alternativos (al modelo de *probabilidad lineal*) para variables *dependientes cualitativas*. Ambos modelos (*logit* y *probit*) se asemejan bastante. Si bien, el *análisis logit* ha alcanzado un mayor desarrollo, en parte debido a su mayor practicabilidad. De ahí que se le conceda un mayor protagonismo en esta breve reseña analítica.

El *análisis logit* constituye un modelo de respuesta cualitativa, que mide la relación entre una serie de variables *independientes (métricas o no métricas)* y una única variable *dependiente categórica (no métrica)*. Para ello se basa en el análisis de la *razón de probabilidad* de una variable *dependiente*, en función de las *independientes*. De lo que se trata es de graduar la *probabilidad* de que determinadas variables *independientes* provoquen la ocurrencia de un evento concreto (la variable *dependiente*).

Del análisis se obtiene una *ecuación* similar a la de *regresión*. Los *coeficientes* (calculados siguiendo el criterio de *máxima verosimilitud*, o el de *mínimos cuadrados*) se interpretan como en *regresión*. La *significatividad* de cada uno de ellos también se comprueba mediante el estadístico "t".

La peculiaridad del análisis reside, no obstante, en el cálculo de los *incrementos de probabilidad*. Éstos se calculan para cada variable con *coeficiente* significativo. Se considera el valor del *coeficiente* (β) y el *valor medio* de la variable en la *muestra* (X_j), siguiendo la expresión de McFadden (1974):

$$\text{Prob}(Y=1) = F(X'_j \beta) = \frac{e^{X'_j \beta}}{1 + e^{X'_j \beta}}$$

La *significatividad* del modelo global se comprueba mediante el estadístico X^2 . Este estadístico se complementa con el *porcentaje de aciertos* (el porcentaje de observaciones que logran ser correctamente estimadas mediante los *coeficientes logit*), como un indicador del éxito del modelo en la predicción de la variable *dependiente*.

- *Análisis discriminante*

Una técnica de clasificación y de asignación de individuos a grupos, a partir del conocimiento previo de sus características.

Su objetivo básico es la estimación de la relación existente entre una serie de variables *independientes* y una única variable *dependiente no métrica (o categórica)*.

El número de *categorías* que incluya la variable *dependiente* determina los grupos formados a partir de la *muestra*. Estos *grupos* deben, previamente, haberse configurado mediante la aplicación de una *técnica multivariable de interdependencia* (como el análisis de *conglomerados* o el *factorial*).

Con el *análisis discriminante* lo que se pretende es la comprobación de si los *grupos* creados por otra técnica analítica son adecuadamente caracterizados por las variables que les definen (de acuerdo con los análisis previos).

Asimismo, se trata de conocer cuál es la combinación de variables (*funciones discriminantes*) que hace máxima la diferenciación entre los grupos. El conocimiento de estas variables ayudará a la predicción de la probabilidad de pertenencia de un individuo concreto a uno de los grupos diferenciados. Ello dependerá de los valores que presente en las variables *independientes* analizadas.

La ejecución del *análisis discriminante* exige, igualmente, el cumplimiento de unos *supuestos* claves, como son:

- a) La *normalidad* de las variables *independientes*. Se recomienda que éstas sean *métricas*. En caso contrario, habría que introducir modificaciones en el análisis.
- b) *Matrices de covarianzas* iguales en cada grupo.
- c) Inexistencia de *colinealidad* entre las variables *independientes*, para evitar que su información sea redundante en la explicación de la variable *dependiente*.
- d) Ha de haber, al menos, 20 casos por cada variable *independiente* introducida en el análisis. Esto contribuye a la *significatividad* estadística del modelo obtenido.

La distancia entre los grupos se mide mediante el estadístico D^2 de Mahalanobis. A éste se añaden los estadísticos *lambda de Wilks*, F y X^2 , en la comprobación de la *significatividad* de las variables *independientes* en la diferenciación entre los grupos.

La contribución de estas variables se cuantifica mediante los *coeficientes de función discriminante estandarizados* (similares a los *coeficientes beta* en *regresión*) y los *factores de carga discriminantes* (los “*discriminant loadings*”). Sobre todo, por estos últimos, debido a su mayor precisión.

Como en el *análisis logit*, en el *discriminante* también se calcula la *razón de aciertos*, como medida del éxito del modelo en la predicción de la variable *dependiente* (la clasificación de los grupos).

- *Análisis multivariable de la varianza*

También conocido como MANOVA. Representa una extensión del análisis univariado de la varianza (ANOVA), mediante el cual se exploran, simultáneamente, las posibles relaciones existentes entre: varias variables *independientes no métricas* (normalmente referidas como *tratamientos*) y dos o más variables *dependientes métricas*. Lo que permite la medición de las *correlaciones* entre las variables *dependientes* y entre las *independientes*.

El uso de este análisis se adecúa, igualmente, a los *diseños experimentales*, en la comprobación de los efectos de distintos *tratamientos*. El fin que se pretende es la determinación de la existencia de diferencias, en los *valores medios* de las variables *dependientes*, en cada *grupo de tratamiento*. Ello exige que:

- a) Las variables *dependientes* se hallen distribuidas *normalmente*.
- b) Se parta de *matrices de varianza-covarianza* iguales en todos los *grupos*. En caso contrario, no podría medirse el efecto específico del *tratamiento* dado (debido a que se partiría de grupos inicialmente diferentes).
- c) El *tamaño muestral* ha de superar al necesario para la *varianza simple*, si se pretende que el modelo adquiera *significatividad* estadística.

La *significatividad* de las diferencias multivariadas entre los *grupos* se comprueba mediante cuatro criterios primordiales: la *raíz máxima de Roy*, la *Lambda de Wilks*, la *traza de Hotelling* y la de *Pillai*. Si estos tests multivariados resultan significativos, se procede a la comprobación de las diferencias grupales en cada una de las variables *dependientes*. Para ello se acude al estadístico “*F*”.

• *Correlación canónica*

Una *técnica de dependencia* que permite la comprobación de la existencia de interrelación entre una serie de variables *dependientes* y otra serie de variables *independientes*. Ambas pueden ser tanto *métricas* como *no métricas*.

La finalidad principal del análisis es la obtención de *combinaciones lineales* de cada serie de variables (*dependientes e independientes*). Estas combinaciones han de maximizar las *correlaciones* entre las variables.

El grado de relación entre la serie de *variables canónicas* se mide mediante el *coeficiente de correlación canónica*. Éste, elevado al cuadrado (R^2), representa el porcentaje de *varianza* de una combinación de variables *dependientes canónicas* que es explicada por una combinación de las variables *independientes*. Su *significatividad* se comprueba, de nuevo, mediante el estadístico “*F*”.

Las contribuciones de cada variable (*dependiente e independiente*) a la combinación o serie de *variables canónicas* respectiva se comprueba, al igual que en el *análisis factorial y discriminante*, mediante los *factores de carga* (“*canonical loadings*”). Estos han de ser $\geq .30$ para que se consideren relevantes. La contribución de las variables también puede comprobarse mediante los *pesos canónicos* (“*canonical weights*”), si bien estos últimos presentan una mayor inestabilidad que los anteriores (los *factores de carga*).

• *Ecuaciones estructurales*

También conocidas como modelos LISREL, en referencia a uno de los programas estadísticos más populares para su resolución.

Esta última *técnica de dependencia* puede catalogarse como una extensión del *análisis factorial* y de *regresión múltiple*, por dos razones fundamentales:

- a) En las relaciones de *dependencia* se representan variables *latentes* o *constructos* (no observadas), a partir de valores conocidos en las variables *manifiestas* (u observadas) –a semejanza con el *análisis factorial*–.
- b) Las diversas *relaciones causales* tienen su expresión matemática en varias *ecuaciones de regresión* (ahora denominadas *ecuaciones lineales estructurales*).

Pero, a diferencia de las *técnicas de dependencia* anteriormente expuestas, en los *modelos de ecuaciones estructurales* se comprueba (simultáneamente) diversas *relaciones causales*. No sólo entre varias variables *independientes* (*métricas o no métricas*) y *dependientes* (*métricas*), sino entre ellas mismas, también (dependientes con dependientes e independientes con independientes). De ello resulta un *modelo causal* de mayor complejidad, aunque más ajustado a la pluridimensionalidad de las *relaciones causales*.

Las *relaciones causales* también se representan gráficamente, mediante un diagrama de “sendero” (“*path*”). En él se diferencian las variables *latentes* de las *observadas*, además de los *errores de medición* (de cada modalidad de variable), y el tipo de relación entre ellas (por ejemplo, una flecha bidireccional indica la existencia de correlación entre dos variables).

Para cada variable que recibe una flecha se formula una *ecuación*. En esa *ecuación*, dicha variable actúa como variable *dependiente*.

La realización de esta técnica analítica exige el cumplimiento de los mismos supuestos que en las anteriores (como la *independencia* de las observaciones; la *linealidad* de las relaciones; la *normalidad* multivariable; la selección *aleatoria* de las unidades muestrales; y un elevado *tamaño muestral*).

Los *parámetros* se obtienen mediante el método de *mínimos cuadrados generalizados*, o por el de *máxima verosimilitud*. Si bien, el primer método muestra imprecisión, conforme desciende el tamaño de la *muestra* y aumenta la complejidad del modelo. La *significatividad* de los *parámetros* se comprueba mediante la “*t*”, además de considerarse el *error* de la estimación.

Como se diferencia entre variables *latentes* y *observadas*, se comprueba la *fiabilidad* ($\geq .70$) y la *validez* ($\geq .50$) de la *medición* de los *constructos teóricos* (*variables latentes*). A ello se suma el *ajuste* global del modelo alcanzado respecto al inicial.

Existen distintos *índices de ajuste*. Entre ellos destacan los siguientes: el estadístico X^2 (aunque ahora interesan los valores bajos, porque expresan un mejor *ajuste* entre el modelo obtenido y el propuesto inicialmente); el *índice GFI*, que oscila entre .00 y 1.00 (cuanto más se aproxime a 1, mejor); los *índices TLI* ($\geq .90$), *NFI* ($\geq .90$), *AGFI* ($\geq .90$), y *AIC* ($\geq .70$), entre otros.

9.4.2. Técnicas multivariadas de interdependencia

A diferencia de las técnicas analíticas anteriores, las de *interdependencia* presentan un menor poder predictivo. Mediante ellas se analiza la existencia de *asociación* o

relación mutua entre varias variables, sin diferenciar entre *dependientes e independientes*.

La Figura 9.2 representa las principales técnicas de interdependencia. La elección entre una u otra responde a los mismos criterios fundamentales mencionados en las *técnicas de dependencia*. Concretamente, el objetivo de la investigación y el nivel de medición de las variables. Ahora la diferencia básica entre las técnicas se establece en función del nivel de *medición* mínimo exigido en las variables para su cumplimentación: *métrico o no métrico*.

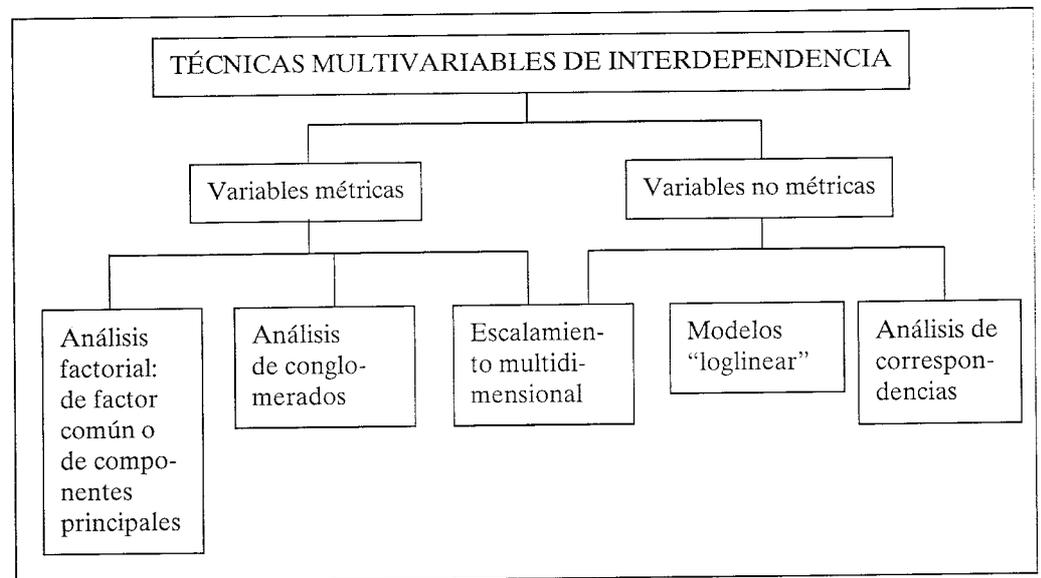


Figura 9.2. Técnicas de análisis multivariante de interdependencia.

- *Análisis factorial*

Una denominación genérica que engloba distintos procedimientos centrados en el análisis de la variación total que una variable comparte con otras variables. Su objetivo fundamental es resumir, la información contenida en un conjunto de variables interrelacionadas, en un número reducido de *dimensiones latentes* comunes (o *factores*).

El *análisis factorial* es *exploratorio*, cuando no se conoce, previamente, cuáles son los "factores". Éstos se determinarán, precisamente, tras el *análisis factorial*. En cambio, el análisis será *confirmatorio*, si se parte de unos "factores" especificados *a priori*, mediante un *análisis factorial exploratorio* u otra técnica multivariante (como el análisis de *conglomerados*). Estos *factores* representan a un conjunto de variables *empíricas* u observadas. Entonces, se tratará de "corroborar" la adecuación de estas variables (empíricas) en la medición de las *dimensiones* de los *conceptos teóricos*.

Además de esta clasificación básica, se diferencia entre análisis factorial de *componentes principales* y de *factor común*. En el *análisis de componentes* se persigue la obtención de *combinaciones lineales* de variables que logren explicar la mayor proporción de *varianza* conjunta. Para ello se tiene en cuenta tanto la *varianza específica* de cada variable, como la *varianza conjunta* (o compartida con otras variables). Por el contrario, en el *análisis de factor común*, la identificación de los *factores latentes* responde sólo a la *varianza común* de una serie de variables.

En cualquiera de las modalidades de *análisis factorial*, el investigador deberá concretar el método para la extracción de los *factores*: *ortogonal* (quartimax, varimax o equimax) u *oblicuo*. En el *ortogonal*, los *factores* se extraen de manera que sean totalmente independientes unos de otros. La extracción *oblicua* permite, en cambio, la existencia de correlación entre los *factores*.

En la decisión de cuántos *factores* escoger han de compaginarse distintos criterios como, por ejemplo, el criterio de *raíz latente* (*autovalor* superior a 1), o el de *porcentaje acumulado de varianza* (al menos superior al 60%).

Las *correlaciones* de las variables *empíricas* con los *factores* se comprueban mediante los *factores de carga* ("*factor loadings*"). Éstos han de ser $\geq .30$ para considerarse significativos.

Una vez definidos los *factores*, se les asigna un nombre o *etiqueta*, que refleje el contenido de las variables *empíricas* que representa.

Las investigaciones comentadas en los Capítulos 4 y 10 ilustran la aplicación de esta técnica analítica. Razón por la que se remite a su lectura.

• *Análisis de conglomerados*

Una variedad de análisis cuya finalidad principal es la clasificación de un grupo de individuos u objetos en un número reducido de grupos. Estos *grupos* han de ser mutuamente excluyentes; han de estar compuestos por individuos lo más similares posible entre sí y diferentes de los integrantes de otros grupos.

A diferencia del *análisis discriminante*, los *grupos* no se hallan predefinidos. Precisamente se aplica esta técnica analítica para la definición de los grupos. Posteriormente, éstos pueden ser corroborados mediante otras técnicas multivariadas (como el *análisis discriminante* o el *factorial confirmatorio*).

El *grupo* se define en función del algoritmo de clasificación que se emplee en la agrupación de los sujetos. En general, se busca la agrupación que haga máxima la distancia entre las *medias grupales* y mínima la *desviación intragrupal* (de los individuos integrantes del grupo respecto a su *centroide* o *media grupal*).

Para la obtención de estos grupos, o *conglomerados*, puede elegirse entre distintos procedimientos, agrupados en dos amplias categorías: *jerárquicos* y *no jerárquicos*. Los primeros se dividen, a su vez, en *aglomerativos* y *disociativos* (o *divisorios*).

- a) El procedimiento *jerárquico aglomerativo* comienza con tantos grupos como individuos en la *muestra*. Paulatinamente van reduciéndose el número de grupos

- y aumentando, en contra, el número de individuos en cada grupo. El programa finaliza cuando un único *conglomerado* agrupa a todos los individuos.
- b) El procedimiento *jerárquico disociativo* (o *divisorio*) procede a la inversa. Parte de un único grupo, que integra a todos los individuos. Poco a poco, este grupo va seccionándose en diferentes subgrupos hasta que, al final, existen tantos grupos como individuos.

En ambos procedimientos *jerárquicos* de formación grupal, la elección del número de *conglomerados* se realiza conforme a una variedad de criterios. Entre ellos destacan: el criterio de la *distancia mínima*; de la *distancia máxima*; el *promedio de las distancias*; el *método Ward*; o el de los *centroides*. Mientras que los primeros criterios consideran la distancia de los individuos, el último tiene en cuenta la distancia entre las *medias (centroides)* de las variables. En cualquiera de estos procedimientos *jerárquicos* los datos pueden visualizarse mediante un *dendograma*: una representación gráfica en forma de árbol.

En los procedimientos de agrupación *no jerárquica* se parte, a diferencia de los anteriores, de una especificación previa de los grupos que desean formarse. De lo que se trata es de encontrar representantes para cada uno de los grupos. Un individuo pertenecerá al grupo cuya distancia a su *centroide* sea menor.

Una vez que los grupos se han constituido (por cualquiera de los procedimientos referidos), se procede a la comparación de las *varianzas* de las variables en relación a los grupos (*intergrupo* e *intragrupo*). Ello permite comprobar si los grupos presentan diferencias en los valores de las variables consideradas en el análisis.

En función del valor del estadístico *F* (en cada una de las variables), se rechazan aquellas variables que no diferencian a los grupos. A partir del valor de la *media* de las variables se extrae, por último, los *rangos* que más separan a los integrantes de cada grupo.

- *Escalamiento multidimensional*

Una variedad analítica análoga al *análisis factorial*. Con él comparte un mismo objetivo: la obtención de un número reducido de *dimensiones*, que permitan caracterizar a determinados objetos o sujetos. Si bien difiere (entre otros aspectos) en el número de *dimensiones* a obtener. Mientras que el *análisis factorial* no impone restricciones al respecto (de hecho el análisis puede efectuarse con un número elevado de *dimensiones* o *factores*), el *escalamiento multidimensional* aconseja su reducción al menor número posible. Ello responde a condicionamientos impuestos para la representación gráfica de los resultados de la investigación.

En el *escalamiento multidimensional* los datos se representan como puntos en un espacio "multidimensional". La distancia habida entre ellos se considera una expresión gráfica de su semejanza o disimilaridad. Esta representación se hace más fac-

tible e interpretable cuando la información puede reducirse a dos o tres *dimensiones*, como máximo.

Para la concreción de estas *dimensiones* se acude a uno (o varios) *criterios de bondad de ajuste*: el “stress” de Kruskal, el “s-stress” de Young, la *correlación múltiple al cuadrado*, o el *diagrama de Shepard*.

Para que la configuración de las *dimensiones* presente un *ajuste* adecuado, los valores correspondientes a “stress” y “s-stress” han de ser bajos. En cambio, la *correlación múltiple al cuadrado* ha de ser elevada. Su valor se interpreta como proporción de *varianza* explicada por las distancias respectivas. Por su parte, el *diagrama de Shepard* ha de reflejar una tendencia ascendente (o creciente).

Dependiendo del nivel de medición de las variables, se diferencia entre *escalamiento multidimensional “métrico”* y “no métrico”.

a) *Métrico*, cuando las variables son de *intervalo* o de *razón*. Lo que favorece su configuración en una escala *continua*.

A partir de una *matriz de correlaciones*, o de *distancias* entre objetos, se trata de situar a éstos en un espacio *multidimensional*. Para ello se transforma la similaridad o disimilaridad percibida en ellos, en distancias (*euclidianas*).

b) *No métrico*, si las variables son *cualitativas (nominales u ordinales)*. En este caso, se parte de una *matriz de rangos*. De ésta se obtiene información de la similaridad de los objetos. La configuración final será aquella que mejor represente a los *rangos* de la *matriz* inicial.

Como en el *análisis factorial*, el *escalamiento multidimensional* puede ser *exploratorio y/o confirmatorio*. Depende de la finalidad del mismo.

• Modelos “log-linear”

También conocidos como modelos “lineales logarítmicos” o modelos “log-lineales”. Su denominación deriva de la transformación logarítmica operada en los datos (las frecuencias observadas en las variables) para facilitar su aditividad: la posibilidad de sumar los distintos efectos de un conjunto de variables *no métricas* (a semejanza de las variables *métricas* en el *análisis de regresión*).

Las *tablas de contingencia* resultan muy útiles en el análisis de variables *cualitativas*. Pero, cuando se añaden terceras o cuartas variables, a modo de variables de *control*, la interpretación de la *tabla* se complica. En estos casos, sobre todo conforme aumenta el número de variables, se aconseja la aplicación de *modelos log-linear*. Éstos se muestran adecuados para el análisis de las interrelaciones entre una serie de variables *no métricas*, que conforman una *tabla de contingencia multidimensional*.

Las casillas de la *tabla* se traducen a componentes, denominados *parámetros lambda*. Estos miden el efecto de los valores de las variables que conforman cada casilla. Si

el valor Z correspondiente a cada λ es ≥ 1.96 , el *parámetro lambda* se considera relevante en la interpretación del modelo. De él se pretende obtener la probabilidad de que un individuo concreto comparta una combinación específica de atributos de un conjunto de variables.

Para la comprobación del *ajuste del modelo* se acude a los estadísticos X^2 y *la razón de verosimilitud*. Los valores de ambos estadísticos tienden a coincidir conforme aumenta el tamaño de la muestra.

- *Análisis de correspondencias*

Una de las *técnicas de interdependencia* de más reciente desarrollo, también adecuada al análisis de variables *cualitativas*.

Parte de la configuración de las variables en una *tabla de contingencia*. Su objetivo fundamental es la representación de las distancias de las *filas* y las *columnas*, que integran la *tabla*, en unos *ejes cartesianos*. Para ello se transforma el valor de la X^2 en una medida métrica de distancia.

Como en el *escalamiento multidimensional*, la proximidad de los puntos mide la similitud existente entre ellos. En el *análisis de correspondencias*, la proximidad muestra, concretamente, la asociación entre las categorías de las variables.

Primero se procede (como en el *análisis factorial* y el *escalamiento multidimensional*) a la identificación del número adecuado de *dimensiones* que categoricen al objeto de estudio. A tal fin, se examina, igualmente, el porcentaje acumulado de *varianza* explicada.

Para cada *dimensión* se deriva, a continuación, unos *autovalores* (“*eigenvalues*”). Éstos expresan la contribución relativa de cada *dimensión* en la explicación de las *varianzas* de las variables.

Una vez establecida la *dimensionalidad*, se identifica la *asociación* o relación existente entre las categorías de las variables mediante su *proximidad*. Ésta ha de comprobarse o en las *filas* o en las *columnas*. Depende de su ubicación. Un valor elevado de X^2 indica un fuerte grado de “correspondencia” entre los atributos de las variables.

De esta forma se obtiene la reducción dimensional de las proporciones de objetos en una serie de atributos. Al mismo tiempo, se extrae la representación de los objetos, relacionados con esos atributos.

9.5. Paquetes estadísticos disponibles

Sin la mediación del ordenador no sería viable ninguno de los análisis estadísticos reseñados. Especialmente, los análisis *multivariados*.

La gran revolución experimentada en los últimos años en el campo de la informática hace que la información que se dé quede obsoleta en un breve período de tiempo.

po. Por esta razón, únicamente van a nombrarse algunos de los principales *paquetes estadísticos*, para conocimiento del lector.

La distinción entre los *paquetes estadísticos* responde a su especificidad. Se diferencia entre programas “genéricos” (que ejecutan la generalidad de los análisis estadísticos), y los “específicos” (especializados en técnicas analíticas concretas).

Entre los *paquetes estadísticos genéricos* destacan el SPSS, BMDP, SAS, ESP, y OSIRIS. En sus manuales respectivos se detallan tanto aspectos técnicos (de funcionamiento del programa), como analíticos (para la interpretación de los análisis estadísticos). Dicha interpretación será la misma, independientemente del programa utilizado.

Los *paquetes estadísticos específicos* son numerosos y variados. A modo de ejemplo se señalan los siguientes: LISREL y EQS (para el análisis de *ecuaciones estructurales*), LIMDEP (análisis *logit*), INDSCAL (*escalamiento multidimensional*), ECTA (*tablas de contingencia*), SPAD (análisis de *correspondencias*), o BROCOLI (*series temporales*).

A estos paquetes estadísticos se suman otros específicos para la realización de *gráficos* (como el SYSTAT, STATGRAPHICS, o el HARVARD GRAPHICS). También hay que mencionar las *bases de datos numéricas* (como DBASE o SIR-DB). Estas *bases* están diseñadas para la exportación automática de datos en la mayoría de los paquetes estadísticos (en especial, los *genéricos*).

Finalmente, cabe mencionar que en los últimos años comienzan a comercializarse *sistemas expertos en análisis de datos* (como el GLIM). Éstos funcionan a modo de consultor experto en *técnicas cuantitativas de análisis*.

Lecturas complementarias

- Bisquerra, R. (1989): *Introducción conceptual al análisis multivariante*. Barcelona, PPU.
- Garrido Luque, A.; J. L. Álvaro Estramiana (1995): *Técnicas de análisis estadístico en ciencias sociales*. Madrid, Universidad Complutense.
- Hair, J. et al. (1992): *Multivariate data analysis*. New York, McMillan.
- Sánchez Carrión, J. J. (comp.) (1984): *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*. Madrid, CIS.
- Sánchez Carrión, J. J. (1989): *Análisis de tablas de contingencia: el uso de los porcentajes en las ciencias sociales*. Madrid, CIS, Monografía n.º 105.
- Sánchez Carrión, J. J. (1995): *Manual de análisis de datos*. Madrid, Alianza Universidad.
- Sánchez Carrión, J. J. y Torcal, M. (1992): *Utilidades del SPSS/PC+. Presentación de informes, grabación de datos y creación de gráficos y mapas*. Madrid, Alianza.
- Spiegel, M. (1991): *Estadística. Teoría y problemas resueltos*. Madrid, McGraw Hill.

Ejercicios Propuestos

1. Interprete los datos que figuran en los Cuadros 9.3. y 9.4.
2. Describa una investigación en la que se haya aplicado alguna técnica analítica multivariable. Especifique los análisis efectuados y su adecuación respecto a los objetivos de la investigación.
3. Trace el plan de análisis que seguiría para obtener el perfil del alumnado de la facultad, añadiendo las razones de su elección.
4. Si la investigación tuviera como objetivo comprobar qué variables influyen en la probabilidad de aprobar una asignatura, ¿qué análisis efectuaría? Justifique la respuesta.
5. A partir de los datos siguientes, confeccione una tabla e interprétela.

<i>Edad</i>	<i>Sexo</i>	<i>Ideología política</i>	<i>Frecuencia absoluta</i>
De 18 a 25 años	Varón	Izquierda	130
		Centro	220
		Derecha	275
	Mujer	Izquierda	115
		Centro	105
		Derecha	310
De 26 a 40 años	Varón	Izquierda	175
		Centro	270
		Derecha	210
	Mujer	Izquierda	196
		Centro	307
		Derecha	150
De 41 a 60 años	Varón	Izquierda	320
		Centro	178
		Derecha	97
	Mujer	Izquierda	205
		Centro	240
		Derecha	170