Nº	FECHA	HORA	DIRECCIÓN	RESULTADO
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7
				1 2 3 4 5 6 7

Anexo 4: Ficha de campo

FICHA DE CAMPO

Nº DE ENCUESTADOR:
POBLACIÓN: ALMERÍA
DISTRITO: 1
SECCIÓN: 3
PUNTO DE MUESTREO: 1
CLAVE PARA LA SELECCIÓN DE VIVIENDA: 35

TABLA DE CUOTAS:

MUJERES										EDAD	HOMBRES									
						■	■	■	■	>60	■	■	■	■						
						■	■	■	■	45-60	■	■	■	■						
				■	■	■	■	■	■	30-44	■	■	■	■	■					
			■	■	■	■	■	■	■	18-29	■	■	■	■	■	■	■			
10	9	8	7	6	5	4	3	2	1		1	2	3	4	5	6	7	8	9	10

CAPÍTULO 8

INTRODUCCIÓN AL TRATAMIENTO DE DATOS

Antonio J. Rojas Tejada
Juan Sebastián Fernández Prados

8.1. Introducción

Empecemos definiendo la palabra *tratamiento* como todo el proceso que recorren los datos desde su etapa de recogida hasta el momento en que se obtienen los resultados de los análisis estadísticos, es decir, justo antes de elaborar las conclusiones sobre ellos. Por *análisis* de datos entendemos la etapa donde los datos ya depurados y codificados se analizan numéricamente y/o gráficamente. El tratamiento de datos, por tanto, es un proceso más amplio que el mero análisis de datos, incluso podríamos considerar este último como una etapa más del primero. En este capítulo nos ocuparemos de las primeras etapas del tratamiento de datos, es decir, de las transformaciones que debemos hacer con los datos hasta que estén listos para llevar a cabo el análisis de los datos. Detenernos en el análisis de datos desbordaría nuestro texto, además de existir en el mercado bastantes que se dedican a ello.

8.2. Tratamiento de datos

El tratamiento de datos comienza en el momento en que tenemos los datos brutos de las encuestas, es decir, las respuestas de todas las personas que han sido encuestadas. Los datos brutos se presentan tal como los hemos recogido y, en nuestro caso, generalmente en forma de cuestionarios, hojas de respuestas, formularios, fichas, anotaciones en papel, etc. Estos datos deben ser *tratados* con el fin de obtener conclusiones a partir de ellos. En primer lugar los datos brutos han de depurarse para evitar los posibles errores que hayamos podido cometer durante la fase de recogida o registro de los mismos. Luego deben codificarse para poder ser tratados de forma numérica o gráfica.

8.2.1. Términos comunes

Antes de seguir adelante, se hace necesario referirnos a algunos de los términos más utilizados en el ámbito del tratamiento de datos en Ciencias Sociales, fundamentalmente

porque es necesario para poder entendernos con todos aquellos que realizan las mismas actividades (Manzano, 1993).

a) Variable: una variable es una representación numérica de una característica (Botella, León y San Martín, 1993), o una propiedad que adopta diferentes valores (Kerlinger, 1987), o cualquier dimensión, atributo o característica que es susceptible de variación (Arnau, 1978). Este concepto se opone al de *constante.*

b) Valor (modalidad): cada uno de los posibles estados con los que puede presentarse una variable se llama valor o modalidad.

c) Sujeto/caso/individuo: de los sujetos/casos/individuos obtenemos las medidas de las variables que nos interesan. En Ciencias Sociales es frecuente utilizar personas, pero también existen otros sujetos/casos/individuos tales como instituciones, organizaciones, etc. Coinciden con las unidades muestrales.

d) Dato: cuando tomamos una medida de una variable de un sujeto caso/individuo obtenemos un dato. Si investigamos a 20 sujetos en tres variables, obtendremos 60 datos. Además, si cada variable puede tomar cinco valores distintos, podemos obtener 60 datos con 300 valores diferentes. Habitualmente, los datos se repetirán, es decir, las variables toman iguales valores en distintos sujetos.

e) Matriz de datos: La matriz de datos está compuesta por los datos que obtenemos al medir las variables en los distintos sujetos. Es decir, nos estamos refiriendo a la matriz de datos de sujetos x variables. Generalmente, los sujetos se sitúan en las filas y las variables en las columnas (cuadro 8.1).

CUADRO 8.1. Matriz de datos

	$Variable_1$	$Variable_2$	$Variable_3$	...	$Variable_m$
$Sujeto_1$	$Dato_{11}$	$Dato_{12}$	$Dato_{13}$	...	$Dato_{1m}$
$Sujeto_2$	$Dato_{21}$	$Dato_{22}$	$Dato_{23}$	...	$Dato_{2m}$
$Sujeto_3$	$Dato_{31}$	$Dato_{32}$	$Dato_{33}$	...	$Dato_{3m}$
.	.	.	.	...	.
.	.	.	.	...	.
$Sujeto_n$	$Dato_{n1}$	$Dato_{n2}$	$Dato_{n3}$	...	$Dato_{nm}$

8.2.2. Codificación de datos

Estamos en el punto en que hemos obtenido las medidas de determinadas variables de una muestra o población de sujetos (datos brutos). Como hemos comentado, esos datos habitualmente se encuentran en fichas, cuestionarios o simplemente sobre un papel. Como podemos imaginar esta información debe ser transformada a datos con los que podamos operar.

En este momento empieza el proceso de codificación. Pero para ser exactos, el proceso de codificación tiene dos partes (Noelle, 1970). Una primera se lleva a cabo en la construcción del cuestionario, ya que gran parte del proceso de codificación viene determinado por las preguntas y, sobre todo, las respuestas que se hayan considerado; de hecho, las categorías o valores de las respuestas al cuestionario serán los códigos de las variables.

La segunda etapa, o codificación propiamente dicha, consiste en llevar a cabo la transformación de las respuestas de los sujetos a códigos o datos que puedan ser operativos. Para hacer este traslado hay que tener en cuenta que los datos que tenemos (datos brutos) pueden proceder de preguntas abiertas o cerradas, y dentro de las cerradas estas pueden ser números o letras. Así vamos a obtener dos tipos de variables: variables alfanuméricas o literales (para las respuestas a preguntas abiertas y cerradas con letras) y variables numéricas (para las respuestas a preguntas cerradas con números).

En general, podemos decir que el proceso de *codificación* consiste en (Etxeberria, Joaristi y Lizasoain, 1991):

1. Nombrar las variables que hemos medido a los sujetos. En el caso de las encuestas las variables suelen coincidir con las preguntas del cuestionario.
2. Asignar códigos a los distintos valores de las variables. La codificación consiste en realizar listas numeradas que contienen todas las posibles respuestas que se dan a cada pregunta. En este sentido, hay que proceder de igual forma ya sea con las preguntas abiertas como con las cerradas. Como veremos más adelante, al proceso de construcción de estas listas numeradas con las respuestas a preguntas abiertas se conoce con el nombre de categorización de respuestas.
   En cualquier caso, es muy aconsejable codificar todos los datos como variables numéricas, debido a que estas variables son mucho más fáciles de manejar para un programa estadístico, de hecho, muchos de estos programas no admiten literales o variables alfanuméricas. Recordemos la importancia que van a tener estos valores numéricos en la tabulación y análisis de datos.
3. Asignar un código específico a los valores ausentes *(missing values),* es decir, asignar un número a aquellos valores que no se han respondido, que son confusos, erróneos, etc. Generalmente se utilizan números que no son valores de las variables, por ejemplo el 9, 99, 999, etc.
4. Construir la matriz de datos. Generalmente, se realiza en papel, para posteriormente grabarla en soporte magnético.
5. Grabarla en soporte magnético. Hay que tener en cuenta que la asignación de números a variables alfanuméricas no garantiza que podamos hacer las operaciones que creamos pertinentes. Para ello, hay que tener en cuenta las escalas de medida (nominal, ordinal, intervalo, razón) en que vienen expresadas las variables (para mayor información sobre las escalas de medida se puede consultar cualquier manual de análisis de datos o psicometría, no obstante, y por citar algún libro que aparece en el apartado de referencias, señalamos que en el texto de Álvaro y Garrido –1995– hay un capítulo dedicado a ellas).

Existen algunas consideraciones a la hora de llevar a cabo este proceso de codificación, debiéndolas tener presentes desde la fase de construcción del cuestionario, entre las que cabe destacar las siguientes (Hague y Jackson, 1994):

— No se deberían emplear más de 10 códigos por variable. Esta consideración es importante ya que con más de 10 códigos es complicado mostrar una imagen general de la variable medida.
— En general, los códigos que reflejen opciones que recojan pocas respuestas (menos del 5%) no son de demasiada utilidad.

— De igual forma, los códigos que reflejen la opción de "otros" no deben recoger más del 10% de las respuestas.

En cuanto a las preguntas abiertas, debemos llevar a cabo un proceso de categorización de todas las respuestas que se han dado. Para ello es necesario leer todas y cada una de las respuestas dadas y establecer, con criterios lógicos y motivados teóricamente, distintas categorías que agrupen todas las respuestas. Como podemos imaginar, la categorización suele estar caracterizada por su tediosa labor y por sus múltiples problemas. De tal forma es así, que la experiencia de no pocos investigadores noveles hace que sólo en sus primeras investigaciones mediante encuestas utilicen preguntas abiertas en sus cuestionarios, para pasar posteriormente a incluirlas cerradas. Esto es así porque el proceso de categorización de las preguntas cerradas implica un gran conocimiento sobre el tema para establecer buenos criterios; además de representar un gran coste a nivel económico, de tiempo y de esfuerzo (Hague y Jackson, 1994). Tras esta categorización, las respuestas a las preguntas abiertas están totalmente acotadas, y, por tanto, el proceso de codificación se hace similar al de las preguntas cerradas.

8.2.3. Formato de los datos

Una vez que tenemos codificadas las variables para cada sujeto, tenemos que ponernos a pensar en cuál será el formato que aplicaremos a dichos datos.
Vamos a distinguir varios tipos de formatos de datos (Manzano, 1993):

1. *Caótico:* es decir, ni nosotros mismos sabemos a qué sujeto corresponde el valor de una determinada variable. Por ejemplo, si tenemos seis sujetos y tres variables, obtendremos 15 datos.

```
223 12 22 6   12 8  654 2 56 1   33 8 6 345 33 3 2 234
```

   Como los lectores habrán podido imaginar, este tipo de formato sólo se utiliza para textos como éste. Nunca se deberían aplicar.

2. *Formato fijo.* En este formato cada variable ocupa una columna y cada individuo una fila. Se llama fijo porque la posición que ocupa cada variable es siempre la misma columna. Las columnas tendrán tantos dígitos como valores hayamos asignado a cada variable.
   Si seguimos con el ejemplo anterior, podríamos obtener la siguiente matriz (llamado también *formato fijo espaciado):*

```
223 12 6
654  6 8
345 33 3
 56 33 8
234 22 2
 12  2 1
```


   O si no queremos dejar espacios entre columnas (llamado también *formato fijo compacto):*

```
223126
654068
345333
056338
234222
012021
```

   Este formato es un poco tedioso, ya que debemos especificar la posición exacta de cada variable. Pero es especialmente útil para la corrección de errores cometidos al introducir los datos.

3. *Formato libre.* Al utilizar este formato cada valor debe estar separado del anterior y del siguiente, al menos, por un espacio en blanco. A diferencia del anterior, ni las variables tienen por qué ocupar la misma columna ni cada sujeto tiene por qué ir en una fila distinta. Veámoslo con el ejemplo.

   Con un sujeto por fila:

```
223 12 6
654 6 8
345 33 3
56 33 8
234 22 2
12 2 1
```

   Con dos sujetos por fila:

```
223 12 6 654 6 8
345 33 3 56 33 8
234 22 2 12 2 1
```

   Los programas que pueden utilizar este tipo de formato leen los datos en el orden en que se encuentran y los van asignando a las variables y a los sujetos que correspondan porque la cantidad de datos leídos así lo determina.
   Este formato tiene la ventaja de ser el más cómodo para escribir los datos, además de no ser necesario indicar posteriormente en qué columnas se encuentra cada variable. Tiene por inconvenientes el de no detectar fácilmente los errores en los datos y que si no existe uno de los valores *(missing)* se debe especificar claramente que ese es un valor *missing* (cosa que no ocurre con formato fijo, ya que basta con dejar el espacio en blanco).

Una vez que conocemos los dos formatos de introducción de datos, la elección depende del especial cariño que uno le tenga a uno u otro. No obstante, existen estructuras más aconsejables que otras en según qué casos (Manzano, 1993):

— Si tenemos muchos datos que introducir, posiblemente no los revisaremos nunca visualmente, sino a través de un software especializado. El formato más aconsejable es el fijo compacto (sin espacios entre variables), pues mantiene un mismo esquema y ocupa el mínimo espacio en el disco.
— Si tenemos pocos datos y los vamos a revisar visualmente, el formato ideal es el fijo espaciado, pues mantiene un mismo esquema (ideal para identificación de variables e individuos) y facilita la localización mediante el espaciado.
— Si los datos que tenemos hemos decidido introducirlos directamente en el programa de análisis de datos que vayamos a utilizar, y no a través de un archivo previo, el mejor formato es el libre, pues las probabilidades de error disminuyen considerablemente y resulta más cómodo (si el programa nos lo permite, claro). La revisión se puede realizar pidiendo un listado de los datos por pantalla.

8.2.4. Escritura de los datos

Tras decidir qué formato tendrán los datos, es el momento de comenzar a escribirlos en el ordenador. Ahora debemos decidir qué programa vamos a utilizar para ello. En cualquier caso, un buen consejo es escribirlos en algún programa de texto que permita grabar archivos en código ASCII, ya que, prácticamente, es el código universal para la transmisión de datos.

Casi todos los programas de análisis de datos tienen un editor donde podemos introducir estos datos. En cualquier caso, debemos estudiar cómo operar con ese editor. Por ejemplo, en los programas SAS, SPSS, BMDP, STATGRAPHICS, etc., se pueden escribir directamente los datos con su editor o módulo de escritura de datos. De hecho, trabajar en estos programas en sus versiones para Windows se ha convertido en la delicia de los codificadores de datos. No obstante, también tenemos la opción, que resulta muy cómoda sobre todo para personas que no estén familiarizados con estos programas estadísticos, de trabajar con un procesador de textos que nos permita grabar los datos en código ASCII (o formato DOS). Tan sólo tenemos que saber cómo nuestro procesador de textos exporta estos datos en código ASCII y cómo importarlos desde el paquete estadístico que estemos utilizando.

Lo que sí parece imprescindible es realizar una *hoja o libro de códigos* donde se exprese claramente el número de variable, la localización en número de dígitos que ocupa en la matriz

CUADRO 8.2. Hoja o libro de códigos.

N.° VARIABLE	LOCALIZACIÓN		NOMBRE VARIABLE	ETIQUETA VARIABLE	VALORES	ETIQUETAS VALORES
	C. Inicio	C. Fin				
1	1	5	ide	número de identificación	00001 a 05351	número sujetos
2	6	6	sex	sexo	0 1	mujer hombre
3	7	7	cls	clase social	0 1 2	alta media baja
4	8	9	eda	edad	00 a 99	n° años

de datos (columna de inicio-fin), el nombre de la variable (cómo la hemos llamado), la etiqueta de la variable (su nombre completo), los valores que puede tomar (números utilizados) y las etiquetas de estos valores (sus nombres completos). Véase el ejemplo del cuadro 8.2.

8.2.5. Errores en los datos

Especial cuidado hay que prestar a los errores en los datos, que se pueden producir por múltiples razones. Queremos hacer énfasis en los errores que se cometen en el proceso de codificación y transcripción. Para ello existen varios procedimientos que pretenden reducir este tipo de errores. Aquí nos referiremos brevemente a la prevención del error y a la depuración de datos (Melià, 1990).

Como medidas *preventivas* destacaríamos la utilización del formato fijo tanto en plantillas de papel formateadas con el nombre de las variable, como en presentaciones en la pantalla del ordenador con el formato de las variables (la práctica totalidad de programas de tratamiento de textos, hojas de cálculo o incluso los de análisis estadísticos lo permiten). También la utilización de variables de control, tales como el número de identificación de sujetos, hace que se eviten errores en la introducción de datos.

La depuración de datos consiste en detectar los posibles errores que puedan tener los datos ya introducidos en el ordenador. Melià (1990) habla de técnicas sistemáticas de detección de datos erróneos y técnicas de comprobación. Las *sistemáticas* consisten en hacer análisis de descriptivos (numéricos o gráficos, generalmente de frecuencias) de las distintas variables, para detectar si existen valores o modalidades no permitidas (fuera de rango, es decir, valores por encima o por debajo de los valores permitidos). La *comprobación* consiste en detectar errores comparando los valores de los sujetos en la matriz de datos codificadas con las respuestas brutas de ese sujeto en el cuestionario. Estas comprobaciones (que nunca suelen ser totales) se realizan sobre una muestra aleatoria que recoja entre el 10% y el 20% de los sujetos.

8.3. Análisis estadísticos de datos

Una vez que tenemos la matriz de datos en soporte magnético y una vez decidido el paquete estadístico que utilizaremos, sólo nos resta ponernos manos a la obra para hacer los análisis estadísticos que estimemos convenientes.

Pero, ¿qué análisis estadísticos realizar para datos procedentes de encuestas? La respuesta, como casi siempre, depende. Al igual que sucede con otras técnicas de investigación, no existen unos análisis específicos de la técnica de encuesta, sino que depende en cada caso de los objetivos de la investigación (y de la naturaleza de los datos con que se trabaja).

Los principales aspectos que determinan la utilización de los análisis estadísticos univariantes o multivariantes son (Gómez, 1990: 263):

a) Tipo de encuesta que pretendemos realizar y número de variables implicadas: si pretendemos describir las características de ciertas variables en la muestra en un momento dado, se aplicarán análisis estadísticos univariantes (encuestas descriptivas); si pretendemos relacionar, analizar o comparar varias características de la muestra entre sí, es más adecuado pensar en técnicas multivariantes (encuestas analíticas o explicativas).

b) Nivel de medida de los datos: cuanto mayor sea el nivel de medida de las variables implicadas en la encuesta, mayor es el abanico de técnicas estadísticas que podemos aplicar, además de ser más potentes y sofisticadas.

Dado que el objetivo del capítulo no es el análisis de datos, remitimos a los lectores al anexo 1, donde citamos algunos de los estadísticos más utilizados en el análisis de datos procedentes de encuestas.

Anexo 1: Principales análisis estadísticos que se aplican a los datos procedentes de encuestas

Tablas de frecuencias unidimensionales	Frecuencia absoluta Frecuencia relativa Frecuencia absoluta acumulada Frecuencia relativa acumulada
Tablas de frecuencias bidimensionales o tablas de contingencia	Frecuencia absoluta Frecuencia relativa Frecuencia absoluta acumulada Frecuencia relativa acumulada
Representaciones gráficas	Diagrama de rectángulos o bloques Diagrama de sectores Pictogramas Cartogramas Diagrama de barras Histograma Polígono de frecuencias Diagrama de dispersión
Medidas de posición centrales	Media Mediana Moda
Medidas de posición no centrales	Cuartiles Deciles Percentiles
Medidas de dispersión absolutas	Amplitud o Rango Rango Intercuartil Varianza Desviación típica
Medidas de dispersión relativas	Coeficiente de variación de Pearson
Medidas de forma	Asimetría Apuntamiento o curtosis
Coeficientes de correlación	Lineal Múltiple Parcial
Regresión	Lineal simple Lineal múltiple
	Logística
Contrastes de hipótesis	t test ANOVA ANCOVA

Si los lectores quieren tener una información más detallada pueden consultar cualquier manual general de análisis de datos. El texto de Fink (1995c) es un libro muy sencillo, claro y breve donde se explican estos conceptos estadísticos con múltiples ejemplos de datos procedentes de encuestas. El texto de Santesmases (1997) es un muy buen texto donde no sólo se comentan los análisis estadísticos más empleados (con un apartado para análisis multivariados) en el contexto de la investigación mediante encuestas, sino que además se implementa en un programa informático.