

---

# Proceso de Datos en Psicología

*Teoría y Práctica*

**Pedro M. Valero Mora**  
**Universitat de València**

---



# INDICE

# Contenidos

<b>Introducción</b>	<b>7</b>
Prefacio	9
Concepto	15
La Importancia del Proceso de Datos	23
La atención dedicada al Proceso de Datos	26
<b>1.0 Definición y Estructura de Datos</b>	<b>31</b>
1.1 Introducción	31
1.2 Definición de datos	34
1.3 Fuentes de datos	36
1.4 Estructuras de datos	37
1.5 Casos especiales	43
1.5.1 Datos textuales	43
1.5.2 Datos de categorías secuenciales	44
1.6 Estructuras de datos y computadores	47
1.6.1 Estructuras de datos con tablas rectangulares	47
1.6.2 Estructura de datos en lenguajes de prog. estadística	51
1.6.3 Estructura de datos avanzada: estructura relacional	54
<b>2.0 Preparación y Codificación de Datos</b>	<b>63</b>
2.1 Introducción	63
2.2 Diseño del cuestionario	65
2.2.1 La forma de la pregunta	67
2.2.2 Preguntas cerradas	67
2.2.3 Preguntas abiertas	69
2.3 Codificación de textos ayudada por ordenador	74
2.4 Diseño del cuestionario para la introducción de datos	78
2.5 Codificación de datos	80
2.5.1 Variables nominales	83
2.5.2 Variables ordinales	85

## Contenidos

2.5.3	Variables de intervalo/razón	85
2.5.4	El registro de la codificación	86
2.5.5	Casos especiales	88
2.5.5.1	Valores faltantes, ausentes o no-respuestas	88
2.5.5.2	Valores imposibles	90
2.5.5.3	Variables de respuesta múltiple	91
2.5.5.4	Codificación para el análisis de variables cat.	93
2.5.5.5	Variables ponderadoras	94
2.5.5.6	Variables selectoras y de grupo	96
<b>3.0</b>	<b>Introducción de datos</b>	<b>99</b>
3.1	Introducción	99
3.2	La información es registrada e introducida manualmente	100
3.2.1	Editores y/o procesadores de texto	101
3.2.2	Hojas de cálculo	103
3.2.3	Bases de datos	104
3.2.4	Programas diseñados para paquetes estadísticos	109
3.3	Introducida automáticamente	112
3.4	Introducida manualmente	114
3.4.1	Cuestionario pasado por el administrador	116
3.4.2	Cuestionarios autoadministrados	120
3.4.3	Consideraciones de diseño de pantallas	123
<b>4.0</b>	<b>Importación/Exportación de datos</b>	<b>127</b>
4.1	Introducción	127
4.2	Formatos	128
4.2.1	Codificación de caracteres	129
4.2.2	Codificación de tablas de datos en formato ASCII	129
4.2.3	Codificación de números	131
4.2.4	Otros formatos más específicos	132

4.2.5	Programas de traducción de datos	133
<b>5.0</b>	<b>Calidad de Datos</b>	<b>135</b>
5.1	Introducción	135
5.2	Control de la fidelidad	136
5.3	Control de la consistencia	137
5.3.1	Pruebas exactas o determinísticas	138
5.3.2	Valores inusuales pero no errores determinísticos	143
5.3.3	La distinción entre outliers y errores de datos	145
5.3.3.1	Métodos univariados: Métodos directos	146
5.3.3.2	Métodos univariados: Métodos estadísticos.	147
5.3.3.3	Estadísticos descriptivos para Outliers univa.	149
5.3.3.4	Métodos gráficos para outliers univariados	151
5.3.3.5	Métodos para Outliers multivariados	155
5.3.3.6	Métodos para outliers multivariados	156
<b>6.0</b>	<b>Valores faltantes</b>	<b>167</b>
6.1	Introducción	167
6.2	Datos completa o parcialmente faltantes	168
6.3	Mecanismos que llevan a valores faltantes	169
6.4	Consecuencias de los valores faltantes	171
6.5	Análisis de casos completos frente a análisis de datos dispo.	174
6.6	Otras Soluciones al problema de los valores faltantes	181
6.7	Exploración de valores faltantes	182
6.7.1	Exploración univariada de datos faltantes	184
6.7.2	Exploración bivariada/multivariada de datos faltantes	185
6.8	Asignación de valores faltantes	192
6.8.1	Métodos rápidos	193
6.8.2	Métodos basados en selección de otros candidatos	195
6.8.3	Métodos combinados	197

## Contenidos

6.8.4	Métodos basados en máxima verosimilitud	197
6.8.5	EM para datos bivariados	199
6.8.6	EM para datos normales multivariados incompletos	201
6.8.7	El operador SWEEP (Barrer)	204
6.8.8	Asignación múltiple	207
6.9	Consideraciones finales	210
<b>7.0</b>	<b>Transformaciones</b>	<b>213</b>
7.1	Manipulación de ficheros	215
7.2	Manipulación de variables	222
7.2.1	Transformaciones que afectan a una variable	222
7.2.2	Generación de variables	224
7.2.3	Transformaciones sobre 2 o más variables	225
7.3	Concepto de reexpresión	226
7.3.1	Tipos de transformaciones	228
7.3.2	La escalera de potencias	232
7.3.3	Otras transformaciones	237
7.3.4	La justificación de las reexpresiones	240
7.4	El registro del proceso de manipulación de ficheros	244
<b>8.0</b>	<b>Programación estadística</b>	<b>247</b>
8.1	Introducción	247
8.2	Comparación entre tipos de lenguajes	252
8.3	Descripción de entornos de programación estadística	255
8.4	Conceptos básicos de programación	259
	<b>Referencias</b>	<b>271</b>





# ***Introducción***

Sanmartín y Algarabel (1990) señalan tres tipos de relaciones entre Informática y Psicología:

- a) Psicología e Informática como disciplinas que comparten una serie de intereses y objetivos.
- b) La Psicología como disciplina instrumental para el desarrollo de aplicaciones informáticas.
- c) La Informática como disciplina instrumental para la investigación y aplicación psicológica.

Con respecto al primer tipo de relaciones es habitual señalar la importancia que muchos conceptos provenientes de la Informática han sido tomados para la Psicología (Norman, 1987; Bajo Delgado y Cañas Molina, 1991; Pylyshyn, 1988; Quiñones et al., 1989) contribuyendo al desarrollo de la Psicología Cognitiva y la Ciencia Cognitiva producido en las últimas décadas.

El segundo tipo de relaciones constituye un área aplicada de la Psicología que ha contribuido al progreso en el uso de ordenadores experimentado en los últimos años debido a las mejoras en su facilidad de uso. Ello ha permitido que cada vez más cantidad

de público sea capaz de acceder al uso de ordenadores y que todo lo relacionado con ellos se esté convirtiendo en parte de la cultura general.

Por último, el tercer tipo de relaciones es el que nos ocupará durante el resto de este texto. Hace referencia al uso de los ordenadores como instrumento de investigación y aplicación psicológica. En concreto, Sanmartín y Algarabel (1990) señalan las siguientes:

- a) Procesamiento y análisis de datos.
- b) Control y aplicación de experimentos y tests.
- c) Simulación de modelos matemáticos.
- d) Almacenamiento y búsqueda selectiva de información.
- e) Sistemas automatizados de evaluación, diagnóstico y enseñanza.

Este texto se centrará en la parte concreta de Procesamiento de los Datos. En el primer apartado daremos una explicación acerca del concepto de esa materia. En segundo lugar, se encuentra la sección dedicada a contenidos la cual presenta una extensión notablemente mayor que el resto intentando de ese modo compensar la falta de textos disponibles acerca de esta materia en nuestro mercado. En tercer lugar trataremos lo referido al método y a las fuentes documentales. Finalmente describiremos el programa docente de la asignatura seguido por las referencias utilizadas.

# ***Prefacio***

## **Introducción**

Es habitual señalar el crecimiento de la importancia de los ordenadores personales en muchos aspectos tanto de la vida cotidiana como profesional acaecido en los últimos años. Naturalmente, algo que reclama de una manera tan natural el uso de ordenadores como es el cálculo estadístico ha sido uno de los aspectos que ha recibido mayor impacto. Mientras que hace aproximadamente algo más de una década resultaba habitual llevar a cabo análisis de datos en *mainframes* que aceptaban los datos en fichas de cartón y que devolvían los resultados impresos al cabo de un día o quizás más, en la actualidad están disponibles ordenadores personales que son capaces de esos mismos análisis de una manera casi instantánea y que pueden captar los datos fácilmente por medio de un lector óptico.

De hecho, estos avances no sólo han sido capaces de mejorar la rapidez con la que se realizan los cálculos estadísticos, sino que este aumento ha permitido poner en práctica técnicas o métodos de análisis que resultarían imposibles de manejar sin toda esa potencia computacional. Por ejemplo, la utilización de métodos iterativos de cálculo para resolver problemas que no tienen una solución directa sencilla es la base de técnicas estadísticas que son hoy por hoy totalmente habituales. También, los métodos basados en simulación como el muestreo repetido (*resampling statistics*) permiten una nueva aproximación a la probabilidad. Finalmente, un avance que no suele ser relacionado con cuestiones

estadísticas en un primer momento son las capacidades gráficas de pantallas y ordenadores que han permitido el desarrollo de nuevas técnicas de visualización basadas en representaciones gráficas interactivas que dan una nueva dimensión a la enseñanza y la práctica de la Estadística.

Todos estos progresos tomados conjuntamente han hecho posible algo que parecía imposible no hace demasiados años. Cualquier individuo con unos conocimientos medios de Estadística y Metodología en general podía por sus propios medios llevar a cabo un estudio de tamaño medio o incluso grande. Y decimos podría porque tal y como se describe en la siguiente carta enviada recientemente a la lista de correo EdStat-l, suele ocurrir generalmente que estos individuos carecen de un fragmento de conocimiento que hace esta tarea mucho más difícil de lo que debería.

From: jpwinter@umsl.edu  
To: edstat-l@jse.stat.ncsu.edu  
Subject: Re: Common textbook faults [long]

My 2 cents:

Yes students have problems learning stats and no textbook is perfect. However, I've noticed a fairly consistent theme among graduate students and faculty that I've worked with: they don't understand basic database management.

In my experience, getting a handle on the data takes as much time and energy as statistically analyzing it. Very bright and educated researchers sometimes don't think about how data from the real world has to fit into a spreadsheet like format. Posts to these lists often come from people who have cases and variables messed up or otherwise can't get a "handle" on their dataset.

People can always request help or hire a consultant when they don't understand how to analyze data. Yet if they failed to assign an id number or overwrote (rather than updated) their main data file, then data is lost.

It always seemed funny to me how use of computers was emphasized, but training was limited to a few lines of syntax at the end of the chapter. The computer training I refer to is not how to use SAS or SPSS, but how to capture data, store it, link it, back it up, restore it, etc....

En esta carta vemos que el autor señala que es muy común que la gestión de datos básica se convierta en un problema entre aquellos que intentan llevar a cabo un estudio que implica análisis estadísticos. Así, en su opinión, disponer los datos de una forma correcta en una hoja de datos es algo que se escapa incluso a "Investigadores muy brillantes y educados (...)".

## Concepto

Nuestro concepto de Proceso de Datos está bien expresado en la definición ofrecida por Bourque y Clarke (1992) en la introducción de su libro "*Processing data: The survey example*".

*"Proceso de Datos se refiere normalmente a la conversión de información verbal o escrita en datos que son leídos por una máquina.*

*Bajo esta definición, Proceso de Datos incluye codificación de datos, introducción de datos codificados en un ordenador, verificación de datos y realización de comprobaciones de consistencia y rango sobre los archivos de datos. No obstante, nosotros preferimos una definición más amplia. Para nosotros el Proceso de Datos comienza con seleccionar una estrategia de datos y acaba cuando las transformaciones de los datos están completas. Esta definición incluye lo siguiente:*

- *Desarrollar categorías de respuesta para preguntas precodificadas o abiertas e incorporar las categorías dentro del instrumento de recogida de los datos.*

- *Recoger los datos.*

- *Crear los archivos de datos que pueden ser usados por los paquetes estadísticos.*

- *Transformar los datos en variables útiles para los análisis.*

- *Documentar todos los aspectos del estudio, incluyendo la racionalidad y las decisiones de codificación específicas, así como las transformaciones realizadas".*

En esta definición hay una serie de elementos que es necesario describir con más detalle a continuación para dar una mayor claridad al concepto del Proceso de Datos.

- **"Proceso o Procesamiento de Datos":** El término inglés *Dataprocesing* parece mejor traducido por la expresión *Procesamiento de Datos*. Sin embargo, el uso habitual en nuestro contexto y que se ha convertido en más común es el de *Proceso de Datos*. En nuestra opinión, ambos términos son aceptables y, aunque podríamos identificar los matices que diferencian una expresión de la otra, el hecho de estar creando un significado particular en castellano del contexto de la disciplina de la Psicología que no corresponde exactamente con las palabras que componen esta etiqueta, hace que, en nuestra opinión, esta distinción no sea fundamental. Por ello, nuestro uso será el del

hábito al que estamos acostumbrados (Proceso de Datos), aunque no consideramos erróneo utilizar el término alternativo de *Procesamiento de los Datos*.

• **"Se refiere normalmente"**: La metáfora que podemos utilizar para referirnos al concepto de Proceso de Datos, y que suponemos es apropiada para muchas otras disciplinas, es la de una mancha de aceite que se extiende en un papel junto a otras. Esa mancha tendría una parte central más oscura y bien definida, pero que, a medida que se extiende va perdiendo densidad. A su vez, cuando empieza a hacer contacto con otras gotas, empieza a confundirse con ellas de tal modo que su frontera se borra y resulta difícil determinar cuando acaba una y empieza la siguiente. De este modo, el concepto de Proceso de Datos aquí explicitado tendrá siempre una parte más central y otras más laterales que se confunden con otras disciplinas. Adelantando un tanto el material que viene a continuación tenemos por ejemplo que en el apartado de transformaciones, algunas de ellas podrían ser incluidas como parte de la disciplina de la Estadística, y otras serían mejor comprendidas desde la Psicometría. Otras de las transformaciones, en cambio, al ser solamente operaciones de gestión de datos que afectan principalmente al formato informático de los datos, y no a sus valores propiamente dichos, en cambio serían más centrales al Proceso de Datos. Como resulta poco práctico no realizar mención de esas otras transformaciones una vez emprendido el tema es necesario por tanto abordar cuestiones que tienen un carácter fronterizo, siempre desde el lado del que partimos, pero reconociendo que existe un cuerpo de conocimientos al otro lado que podría incorporar su entramado teórico a continuación.

• **"Convertir información verbal o escrita"**: En muchas disciplinas y en Ciencias Sociales en particular una de las fuentes o canales de transmisión de los datos es el verbal o escrito. En una forma muy habitual de recoger datos, los sujetos simplemente contestan a preguntas o rellenan cuestionarios. Estos datos así recogidos deben ser convertidos antes de empezar la fase del análisis en la que el investigador podrá empezar a extraer las conclusiones en las que esté interesado. Es cierto que también hay situaciones en Ciencias Sociales en las que las variables de interés pueden ser medidas sin utilizar este canal, por ejemplo, cuando se miden tiempos de reacción utilizando un aparato diseñado para tal fin. Sin embargo, en la práctica, los científicos sociales, y los psicólogos dentro de ellos, se encuentran muy a menudo en la práctica ante la situación de que las respuestas obtenidas deben pasar por registros escritos o verbales que necesitan de algún tipo de procesamiento antes de poder ser utilizados. Por ejemplo, un sujeto contestará a una pregunta acerca de su tendencia política de un modo verbal y esto será registrado bien por un entrevistador o bien por él mismo sobre un cuestionario. Puesto que el interés de los científicos sociales es llevar a cabo generalizaciones e inferencias acerca de las actitudes, cogniciones o conductas de los individuos y no de uno sólo,

normalmente la pregunta anterior se hará a un grupo de ellos. La teoría de muestreo nos proporcionará las claves para determinar cuántos sujetos y en qué condiciones deberán ser entrevistados para que las generalizaciones que deseamos hacer tengan la suficiente credibilidad. Esto nos llevará a recoger un bloque de cuestionarios, generalmente con más de una pregunta para cada uno de ellos, sobre los que una vez analizados, queremos basar nuestras conclusiones. Ahora bien, la información dispuesta en cuestionarios resulta muy difícil de utilizar, entender y/o visualizar. Una de las empresas en que los científicos están embarcados más habitualmente es la de ser capaces de visualizar situaciones o procesos que no son observables con claridad de modo habitual. Para lograr este objetivo, llevan a cabo manipulaciones de la situación que hacen por ejemplo que el tiempo transcurra más rápidamente, lo pequeño se vea a un tamaño mayor, o lo excesivamente grande sea reducido a un tamaño abarcable por el ojo y la mente humana. De este modo, la información recogida acerca de sujetos humanos debe ser transformada a un formato que el investigador pueda explorar con comodidad y así permitirle aprehender y exponer con claridad las conclusiones de sus estudios. Ese formato debe permitir esa extracción de conclusiones, pero también debe de ser lo suficientemente flexible como para, en caso de ser necesario, admitir cambios de representación con facilidad que respondan a nuevos planteamientos surgidos en el proceso de indagación científica. Es decir, deseamos tener los datos en un formato que tenga la menor "viscosidad", entendida ésta como la resistencia a adoptar nuevas formas.

El ordenador hoy por hoy es la herramienta que nos permite llevar a cabo esas tareas de representación de la forma más flexible y adecuada. Por ello, la conversión de la información será entendida en la mayoría de los casos como dirigida a convertir los datos a un formato legible por él, lo cual, una vez conseguido, permitirá al científico llevar a cabo su actividad.

Asumamos por un momento que no disponemos de ordenadores y que el Proceso de Datos transforma los cuestionarios en un formato no basado en éstos. Por ejemplo, podríamos organizarlos en una "sabana de datos" basada en papel, que mostrara en filas y columnas (una fila por sujeto y una columna por pregunta o variable). El resultado final nos permitirá contestar a algunas de las preguntas que estamos interesados de una manera más cómoda y eficiente que si utilizáramos los propios cuestionarios de los que hemos tomado los datos. Por ejemplo, esa disposición hace simple obtener medias o proporciones, así como indicaciones de relación entre preguntas (es decir entre columnas en nuestra nueva representación). Obviamente, esta hoja de datos no es un gran avance comparado con el que haríamos si utilizáramos un ordenador ya que el número de operaciones que facilita es mucho menor, pero sin embargo, creemos que ilustra el concepto de conversión de formato que subyace a la disciplina de Proceso de Datos.

• **" Datos manejables por una máquina":** Como hemos señalado anteriormente, la máquina más importante a la que nos referimos aquí es el ordenador. Vemos pues que la definición ofrecida por Bourque y Clark da como objetivo último que los datos puedan ser manejados por éste.

La importancia de este objetivo estriba en las posibilidades de manipulación y extracción de resultados que ofrecen los ordenadores frente a los datos en su formato puro u otros mucho menos convenientes, tal y como la sábana de datos. Por medio de esta máquina resulta posible contestar a preguntas de todo tipo de una manera rápida, permitiendo al investigador aprovechar al máximo los datos de que dispone. Sin ordenadores, muchas de las posibles preguntas que podrían interesar el investigador podrían quedar sin contestación ya que los métodos a aplicar serían tan costosos que obtener los cálculos o las representaciones necesitaría tanto tiempo que las respuestas podrían perder su significado o interés. Esto es más claro cuando se tratan los problemas de tipo social o psicológico, los cuales a menudo están ligados a momentos temporales concretos y su estudio necesita de respuestas obtenidas dentro de unos márgenes. Por ejemplo, un estudio de intención de voto debe realizarse antes de que se produzcan las elecciones correspondientes, pero por otro lado no tiene sentido llevarlo a cabo con mucho tiempo de antelación. Siendo así, los datos deben ser analizados con la mayor rapidez, y el método más rápido para lograr esa velocidad es utilizar un ordenador.

Ahora bien, podemos plantear que esta necesidad no es verdaderamente tan grande cuando se trata de responder a preguntas simples. Es muy posible que el cálculo de proporciones simples o porcentajes no requiera unas operaciones tan complejas como para que no puedan ser ejecutadas por otros medios. Podríamos por ejemplo pedir a los encuestadores que hagan las sumas y las divisiones por sí mismos y que proporcionaran los resultados ya directamente. Dejando aparte el problema de la precisión de los cálculos así obtenidos, este modo de proceder supondría una forma de actuar poco conveniente por las siguientes razones:

- a) El enfoque moderno del Análisis de Datos (Tukey, 1977; Hoaglin, Mosteller y Tukey, 1983; Chambers et al., 1983) considera como parte intrínseca de un análisis de datos el proceso de indagación que, guiado por una serie de preguntas iniciales, enriquece las hipótesis previas, esclareciendo aspectos indeterminados en un principio así como descubriendo fenómenos de interés que merecen una atención especial cuidadosa. Así, el Análisis de Datos adquiere un carácter no planeado rígidamente sino que a menudo requiere una serie de operaciones que se ajustan a lo descubierto en cada momento y que van sugiriendo nuevos rumbos de acción, nuevas pruebas o incluso la obtención de datos adicionales.



Este proceso, que vemos resulta altamente impredecible, necesita de una estrategia de análisis que permita realizar comprobaciones sucesivas de una manera ordenada y rápida. Para ello, la utilización de ordenadores constituye prácticamente la única forma conocida de adaptarse a este contexto cambiante de análisis, al permitir manipulaciones de los datos y la utilización de técnicas de todo tipo. Obviamente, existen todavía muchos aspectos que pueden mejorarse y todos conocemos situaciones en las que resulta ridículamente difícil aprovecharse de las capacidades de dos posibles métodos alternativos de análisis simplemente por las dificultades de traducir la información entre el *software* particular encargado de realizar las operaciones correspondientes.

b) Existen gran cantidad de cálculos estadísticos que, debido a su complejidad, no pueden ser concebidos sin ordenadores. Estos son fundamentalmente los basados en procesos iterativos, los cuales han demostrado un gran valor a la hora de resolver problemas de cálculo que no es interesante resolver de modo explícito (debido a que no se conoce la solución correcta o a que resulta excesivamente complejo derivar el procedimiento matemático) y se prefiere obtener una solución de tipo implícito que, por aproximaciones sucesivas, optimicen una función de ajuste dada. Algunos métodos de análisis estadísticos que estarían dentro de esta categoría son el escalamiento multidimensional y los procedimientos de estimación basados en máxima verosimilitud.

La importancia del ordenador es de gran ayuda también para la comprensión de estos cálculos puesto que conceptos que le son propios tal y como el de bucle o recursión son la base de aquellos, y su visualización facilita enormemente la representación mental de su funcionamiento.

c) La capacidad gráfica de los ordenadores ha supuesto igualmente una modificación en las técnicas de análisis estadístico tradicionales al permitir la utilización de gráficas estadísticas de una manera sencilla y eficaz (Chambers, 1983; Hoaglin, Mosteller y Tukey; 1991; Cleveland y MacGill, 1988).

Quizás el efecto de mayor impacto ha sido el que se ha producido dentro del contexto del Modelo General Lineal. Este modelo de análisis se caracteriza por estar basado en una serie de supuestos que, en caso de no cumplirse, limitan seriamente su aplicabilidad. Por ello, los estadísticos han desarrollado un gran número de pruebas que se dirigen a diagnosticar aspectos como la linealidad de las relaciones, la homoscedasticidad del error y la normalidad de las variables. Estas pruebas estadísticas generalmente eran del tipo de proporcionar un valor que implicaría

aceptar o rechazar el cumplimiento de los supuestos. En caso de producirse, los verdaderos análisis seguirían adelante. En caso de no producirse, en cambio, los análisis no deberían realizarse y el investigador por lo tanto debería seguir un curso alternativo de acción. En algunos casos, la solución podría venir por la utilización de técnicas que utilizaran supuestos menos restrictivos. No obstante, puesto que el Modelo Lineal General se encuentra enormemente desarrollado y permite análisis mucho más sofisticados que esas alternativas menos restrictivas el analista en muchas ocasiones desea no tener que renunciar a su utilización.

Además, en muchos casos, el incumplimiento de los supuestos, a pesar del valor que la prueba encaminada a comprobar su incumplimiento haya proporcionado, no siempre ocurre de una manera absoluta. Muy a menudo resulta posible identificar el caso o casos que tienen un comportamiento extraño y que hace que los resultados aparezcan como negativos. Hoy en día, la mayoría de los libros de texto aconsejan examinar los gráficos estadísticos como un complemento e incluso como un sustituto de estas pruebas. Estos gráficos proporcionan además la oportunidad de obtener una impresión acerca de muchos aspectos simultáneamente, así como de la interacción entre ellos. Por ejemplo, la falta de normalidad en una variable se puede deber exclusivamente a la existencia de valores extremos. Existen técnicas estadísticas que permiten comprobar ambos supuestos pero un gráfico por otro lado dará información acerca de ambos aspectos simultáneamente, mostrando que la falta de normalidad se debe como decíamos a unos cuantos valores extremos.

Podría argumentarse que no es necesario realizar los gráficos estadísticos por medio de un ordenador. Después de todo, el libro de Tukey (1977) está lleno de ejemplos ingeniosos en los que se obtienen representaciones gráficas utilizando solamente herramientas de dibujo muy simple. Después de todo los gráficos estadísticos no necesitan ser una obra de arte, y en la práctica, esos excesos son probablemente contraproducentes (Tufte, 1983).

Un avance relacionado con el tema de los gráficos estadísticos que sí que llama inevitablemente al uso de ordenadores es el de los gráficos dinámicos (Tierney, 1990; Cleveland y MacGill, 1988). En ellos, la animación y la interconexión entre gráficos y datos constituyen una herramienta esencial del análisis, permitiendo superar las limitaciones que ofrece el tener que consultar aspectos parciales de los resultados sin poderlos combinar con coherencia. Por ejemplo, un valor que parece destacar sobre el resto en un aspecto podría destacar en otros aspectos igualmente. Sin embargo, no existe una manera sencilla de representar los diferentes aspectos en un único gráfico. La interactividad entre gráficos supone una forma de superar esta

barrera de una manera simple ya que seleccionando uno de ellos el otro también se seleccionaría, por lo que la identificación sería inmediata. Un ordenador puede gestionar de una manera simple y rápida este tipo de conexiones, haciendo posible el tenerlas en cuenta cuando son observadas. Otros métodos más lentos harían imposible que los analistas realizaran estas comprobaciones al ser excesivamente costosas.

Otra ventaja de los gráficos estadísticos es que su uso puede hacer más sencilla la comprensión y la interpretación de resultados. Podemos comparar por ejemplo un biplot (Gabriel, 1986) con el resultado basado en output de un análisis de componentes principales. El primero ofrece una interpretación mucho más rápida que la basada en el output textual, la cual es más costosa y lenta de evaluar. Estos gráficos pueden ser de gran ayuda para la enseñanza de la Estadística al permitir dar una imagen que haga más fácil la comprensión intuitiva del significado de las técnicas (Young y Bann, 1997).

d) Una última ventaja del uso de ordenadores es la referida a la enseñanza de la Estadística. A menudo surge la polémica entre los partidarios de explicar una "Estadística sin fórmulas" en la que los paquetes estadísticos realizan todos los cálculos y producen los resultados y los usuarios sólo se ocupan de las tareas de Proceso de Datos, ejecución de las rutinas ya elaboradas y posterior interpretación de los resultados, frente a los que opinan que lo importante es concentrarse en la comprensión de los desarrollos matemáticos subyacentes y la aplicación del ordenador debe dejarse en un lejano segundo plano (Valero, Molina y Sanmartín, 1992). Esta es, obviamente, una polémica falsa en cuanto que la mayoría de las personas admitirán que ambos extremos son deseables y, por tanto, sería necesario transmitir algo de ambos a los estudiantes. Sin embargo, si tenemos en cuenta que el tiempo disponible para estas asignaturas está limitado, admitiremos que es necesario un cierto grado de compromiso entre los extremos. En este caso, los ordenadores suponen una modificación de los contenidos a enseñar ya que, por ejemplo, aquellas fórmulas que pueden ser consideradas como abreviaciones que simplemente facilitan el cálculo manual podrían ser evitadas (por ejemplo, la distinción entre la fórmula de la correlación de Pearson, la biserial-puntual y la phi). También, aunque el cálculo manual puede ser útil para comprender mejor ciertos conceptos estadísticos, no es verdaderamente necesario realizar una gran cantidad de repeticiones de éstos con el objetivo de lograr habilidad y precisión con ellos. Así, puede ser más práctico aprender un lenguaje de programación que permita, una vez comprendido un proceso de cálculo, automatizarlo de tal modo que sea el ordenador el que repita la secuencia necesaria para conseguir el resultado. También,

la posibilidad de pensar en los análisis de una manera más global, sin tener que atender a los detalles permite llevar a cabo una experimentación con las técnicas que, sin duda, debe redundar en una mejor comprensión de éstas. Por ejemplo, examinar el resultado de un análisis de regresión con y sin un valor extremo en el espacio de los predictores puede ayudar a entender el concepto de influencia y de valores extremos sobre la aplicación de análisis de regresión. Este tipo de pruebas resultan inmediatas y obvias utilizando un paquete estadístico, y además, si tenemos en cuenta la naturaleza del ordenador, con la ventaja de ser directamente comparables entre sí, así que el estudiante puede estar seguro de que la diferencia que observa en, digamos, los coeficientes de la regresión pueden ser atribuidos al valor modificado y no a un error de cálculo.

Recapitulando, hemos visto una explicación de los diferentes aspectos incluidos dentro de la definición estricta de Proceso de Datos tal y como es proporcionada por Bourke y Clark (1992). Dentro de esta definición más reducida ellos sugieren que se incluyan los siguientes aspectos:

- Codificación de datos: Puesto que las máquinas utilizan la información según unos formatos diferentes a los verbales o escritos es necesario hacer una traducción que la ponga en correspondencia con aquéllas. Esta codificación debe tener en cuenta por ejemplo los tipos de variables que manejan los ordenadores (p.e. texto, números enteros y números en precisión simple o sencilla) y los datos originales. Por ejemplo, la variable color del pelo podría ser recogida como Rubio, Moreno o Castaño en la vida real pero al ser codificada para ser usada en un ordenador podría pasar fácilmente a utilizar el conjunto de códigos {1, 2, 3}.

- Introducción de datos: Una vez codificados los datos nos encontramos ante la necesidad de traspasar la barrera que los separa de convertirlos en registros electrónicos. Esta es una barrera costosa en cuanto a tiempo y que es muy susceptible de dar lugar a errores o equivocaciones. En la actualidad existen disponibles muchos métodos para llevar a cabo esta tarea que permiten mejorar en gran medida este proceso y evitan una excesiva contaminación por parte del error de datos, lo cual podría llevar a invalidar las conclusiones de nuestro trabajo.

- Verificación de datos: Una vez introducidos los datos en el ordenador, un paso que se considera necesario es llevar a cabo una evaluación que determine hasta qué punto el proceso que éstos han seguido ha terminado con un resultado aceptable, listo para empezar los análisis correspondientes. Estas comprobaciones pueden dar como resultado el descubrir que nuestros datos no son aceptables y que tengamos que reelaborar parte de

él si descubrimos que la información que hemos recogido y luego introducido no resulta aceptable.

Bourque y Clark (1992) señalan dos tipos de comprobaciones a realizar. La comprobación de rangos y la determinación de la consistencia. La primera de ellas se dirige a examinar si los valores introducidos están dentro de los valores admisibles o, por el contrario, nos encontramos con valores fuera de ellos. Por ejemplo, un sujeto no podría haber visto la televisión más de veinticuatro horas al día o no puede obtener una calificación de once en un examen cuyo máximo es de diez.

En las comprobaciones de consistencia se determina si la combinación de dos o más valores resulta admisible desde un punto de vista lógico. Por ejemplo, un hombre no puede haber tenido más de cero embarazos, y un hijo no puede tener más edad que sus padres. Estas comprobaciones son relativamente complicadas pero necesarias cuando la veracidad del dato es central al objetivo del trabajo realizado. Un problema complejo es el que se deriva de tener que determinar qué valor es el incorrecto cuando se detecta una inconsistencia. Si es posible consultar una fuente de datos que resuelva esta cuestión no hay problema naturalmente, pero en caso contrario la estrategia seguramente tendrá que pasar por examinar el resto de valores en la base de datos y aprovechar esta información para estimar el valor correcto. Esta estrategia puede ser considerada una forma de asignación de datos la cual tiene su máximo desarrollo en relación con el tema de los valores ausentes o faltantes.

Un aspecto diferente al de la consistencia lógica pero relacionado con él es el diagnóstico de valores extraños. En muchas ocasiones no existen reglas *a priori* acerca de qué combinación de valores es imposible que se produzca, pero sin embargo si que se pueden observar algunas que podríamos considerar anormales. Por ejemplo, el valor de peso y de altura de los individuos está correlacionado naturalmente. Aunque nada impide que un sujeto pese más de 100 kgs., si tenemos que la altura de éste no supera cierto valor podríamos pensar que un peso de tal calibre se debe a un error de datos.

Los tres conceptos anteriores, no obstante, corresponden a una versión reducida del Proceso de Datos para Bourque y Clark (1992). Ellos mantienen una conceptualización más amplia la cual añade los aspectos que comentaremos a continuación

- Transformar los datos a formato adecuado para llevar a cabo los análisis. Para Bourque y Clark, como para nosotros mismos, el Proceso de los Datos termina con un archivo de datos completamente preparado para su análisis. Para ello es necesario a

menudo realizar transformaciones y preparaciones de los datos que los dispongan correctamente. Ahora bien, mientras que muchas de estas transformaciones pueden ser consideradas como operaciones de gestión de datos y pueden ser descritas por sí mismas, existen otras que se aproximan a la frontera con el Análisis de Datos. Sin embargo, en líneas generales, la forma de realizar eficientemente estas transformaciones es una cuestión que implica cuestiones relacionadas con el uso de bases de datos y las operaciones sobre los registros que les son propias: división, unión, intersección, etc.

Si examinamos los comandos de los paquetes estadísticos veremos que una buena parte de éstos se refieren a operaciones que no son propiamente de Análisis de Datos. Muchos de ellos se refieren a transformaciones de variables, selección de casos, muestreo, unión de ficheros, importación-exportación de datos, etc. Estas operaciones se encuentran separadas en cuanto a organización de las referidas a Análisis de Datos, ya sea con capítulos diferentes en los manuales, con menús diferentes, etc. Además, estas cuestiones son las que están más ligadas directamente a sistemas operativos, siendo el resto similar a través de diferentes plataformas informáticas en las que pueden estar implementados.

• Documentación de los aspectos del estudio. Freedland y Carney (1992) señalan que existen muchas prácticas de gestión de datos que suponen una seria amenaza para la validez de la investigación científica. En concreto, ellos preguntaron a investigadores acerca de la disponibilidad de los datos referente a resultados ya publicados y descubrieron que era muy difícil que los investigadores pudieran proporcionar copias de ellos, debido principalmente a la falta de unas buenas prácticas de documentación. Algunos ejemplos señalados por ellos son los siguientes:

- a) Cuando se realiza un muestreo para llevar a cabo un análisis sería necesario guardar con precisión qué casos fueron seleccionados de la base de datos o los resultados no pueden ser replicados.
- b) Cuando se realizan diferentes copias de los archivos que son analizados y sujetos a transformaciones o modificaciones puede resultar muy difícil identificar exactamente qué archivo es el que se utilizó, para así poder, por ejemplo, replicar unos resultados.
- c) Cambios en el software o renovación de los equipos puede llevar a la pérdida o destrucción de unos datos. También, aunque los datos pueden ser utilizados, puesto que la explicación acerca de a qué se refieren tiene que ser guardada, en muchos sistemas en un lugar separado al que se utiliza para los propios datos, es fácil

encontrar que no es posible saber a qué se refieren los ficheros que tenemos ante nosotros.

Los problemas que conlleva la gestión de los datos ha hecho que una comisión internacional haya desarrollado un lenguaje de documentación de datos denominado Standard Generalized Markup Language-Lenguaje de Etiquetado Generalizado (ICPSR, 1997).

En nuestra opinión, existen todavía otros aspectos en el Procesamiento de Datos que, aunque implícitos en la definición ofrecida por Bourque y Clark (1992), merecen también ser comentados de manera explícita, ya que constituyen una parte fundamental de esta disciplina. Estos serían:

- Análisis de la estructura de los datos. En general, la estructura de los datos que estamos acostumbrados en relación con el trabajo estadístico es el de tabla bidimensional. Esta tabla corresponde en términos matemáticos a una matriz, y en los paquetes estadísticos suele ser la forma más corriente de mostrar nuestros datos puros. Sin embargo, en ocasiones, las estructuras de los datos son mucho más complejas y su conocimiento es un prerequisite a la tarea de inquisición analítica, al limitar o dar forma al tipo de cuestiones que es posible plantear. Por ejemplo, en muchas instituciones los datos están organizados según una estructura relacional en la que determinada información depende de la unidad administrativa u organizativa que los producen o gestionan. Esta información puede estar relacionada con nuestros objetivos de una manera más o menos simple. Así, es muy común que exista una diferencia de nivel o jerarquía entre lo que queremos evaluar y parte de los datos que manejamos. Esta estructuración nos obligará a realizar manipulaciones de los datos que dispongan la información de la manera más adecuada. Todo ello sin olvidar que los paquetes estadísticos suelen exigir una disposición particular, por lo que estas manipulaciones deberán tener en cuenta también lo referido a los análisis que queremos realizar.

- Evaluación de datos faltantes (*missing*). Este aspecto podría ser considerado como un tema situado en un punto medio entre lo referido a la evaluación de la calidad de los datos y las transformaciones. Así, las no-respuestas son un problema de calidad de los datos porque un número o forma de producirse inadecuados producen que, luego, los análisis que es posible realizar sean mucho más limitados. Ello se debe a que la mayoría de los procedimientos estadísticos están concebidos o asumen que los datos están completos para poder ser ejecutados. Cuando esto no se produce, algunos paquetes estadísticos simplemente borrarán internamente aquellos casos que tengan no-respuestas

y producirán un resultado basado en la parte completa (Little y Rubin, 1987). Esto implica una pérdida de información que puede limitar seriamente las conclusiones a extraer tanto por la reducción de la muestra producida como por el posible sesgo incorporado por las peculiares características de los casos con no-respuestas.

- Automatización del Proceso de Datos. Cualquier tarea con el ordenador que implique una secuencia de acciones repetitivas, aunque con posibles caminos divergentes en ciertos puntos, es susceptible de ser automatizada, obteniendo con ello una gran cantidad de ventajas. La programación de ordenadores permite, por un lado, garantizar que los pasos a seguir en cada caso serán repetidos fielmente, sin necesidad de que el operador atienda a detalles parciales de poco alcance. De este modo, puede concentrarse en dar solución a problemas de otra índole. Por otro lado, cuando se producen situaciones especiales, la programación puede incorporar soluciones que incrementan la potencialidad de los programas, dando lugar a un aumento de sus capacidades, del mismo modo que la ciencia contribuye al progreso del conocimiento humano.

Este proceso debería seguir una serie de pasos entre los que tenemos la de definición del algoritmo de solución del problema, su implementación, que obliga a describir con mayor precisión los detalles a tener en cuenta, y finalmente, su depuración, que permitirá investigar la corrección y su comportamiento ante situaciones imprevistas, poco usuales, pero a menudo importantes. Estos problemas a menudo pueden ser resueltos mediante modificaciones del código original, pero, cuando estos son verdaderamente importantes puede ser necesario plantear el proceso desde el principio empezando por un nuevo algoritmo y una nueva implementación.

## La Importancia del Proceso de Datos

Una pregunta de gran importancia es la referida a las situaciones en las que las cuestiones referidas al Proceso de Datos cobran mayor relevancia. Podemos identificar los siguientes aspectos: a) La distancia entre la recogida de los datos y el pasar a estar en el formato de una "máquina" (el ordenador), b) la estructura inicial de los datos, c) el tamaño del problema a manejar y d) el contexto en que aparecen los datos. Veremos estos aspectos a continuación uno por uno.

a) La distancia entre el momento o situación en que son recogidos los datos y su paso a un formato manejable por ordenador. Tal y como veíamos, la definición de Bourque y Clark (1992) asume que los datos recogidos necesitan pasar por una serie de operaciones antes de estar dispuestos correctamente para ser analizados. Estos pasos, dependiendo del contexto particular que consideremos, pueden ser en algunos casos muy



simples y directos, y en este caso, los problemas que resolvería el Proceso de Datos no serían muy complicados. Por ejemplo, cierto estudio podría recoger las contestaciones de los sujetos directamente por medio del ordenador tras la presentación de un estímulo. Los datos luego podrían ser analizados directamente por el propio programa e informados al investigador. Asumiendo que no se van a producir problemas de calidad de los datos, y que el programa que recoge la información está diseñado para que el investigador no necesite ningún esfuerzo, entonces quizás las consideraciones hechas desde el punto de vista del Proceso de Datos son relativamente irrelevantes. Sin embargo, un estudio que implique la recogida de datos a través de un cuestionario que ha sido pasado en diferentes áreas geográficas, que necesita codificaciones, introducción de datos manual en, quizás, diferentes formatos informáticos y, que, por tanto, necesitará de controles de calidad estrictos, reclama de una manera más natural los conocimientos incluidos dentro de la disciplina de Proceso de Datos. De hecho, no es extraño descubrir que los libros dedicados al tema tal y como "Principles of statistical data handling" (Davison, 1996) o "Processing Data: The Survey example" se centren sobre todo en la situación en que los datos son recogidos por medio de cuestionarios en encuestas.

b) La estructura inicial de los datos puede ser un factor de gran importancia. Por ejemplo, muy a menudo, los datos son recogidos a un nivel de agregación diferente al que son analizados, tal y como cuando se recoge la información para unos alumnos en un país, pero los análisis quieren realizarse al nivel de las instituciones, áreas geográficas, u otro tipo de agrupaciones. En este caso, es necesario combinar los resultados individuales al nivel que se quieren extraer las conclusiones. A veces la estructura es más complicada y la jerarquía descrita debería también ser considerada en nuestros análisis, por lo que la gestión de los datos debe poner atención a ella y respetarla adecuadamente. En otros casos, incluso, un modelo jerárquico no será suficiente y los datos se encontrarán dispuestos siguiendo un modelo relacional. Este modelo (Willitts, 1992) resulta el más habitual desde un punto de vista informático, ya que incluye dentro de sí buena parte de las estructuras de datos que podemos concebir. Sin embargo, las técnicas estadísticas disponibles en la actualidad en muy pocos casos han sido diseñadas para tratar esta estructura directamente, por lo que es necesario hacer conversiones que recuperen tablas bidimensionales para llevar a cabo los análisis.

c) El tamaño del problema. Antes de disponer de ordenadores, los investigadores se encontraban con que la fase de cálculo de un estudio podría requerir una gran cantidad de esfuerzo. Ello limitaba el tipo de problemas que era posible estudiar y las técnicas de análisis a utilizar. En la actualidad, los ordenadores han cambiado este panorama y ahora es posible que un investigador no tema enfrentarse a problemas de tamaño relativamente grande en lo que respecta al cálculo. La dificultad, hoy en día, proviene del

Procesamiento de los Datos. Recoger una gran cantidad de cuestionarios, introducirlos de una manera eficiente y controlada y, en definitiva, disponerlos para ser utilizados constituye un problema de mayor envergadura. Este problema tiene, por un lado, una vertiente económica, ya que siempre es necesario ceñirse a un presupuesto que nunca es suficiente para satisfacer todos nuestros objetivos, y, por el otro, sustancial, ya que una gestión deficiente de los datos no detectada a tiempo puede suponer una amenaza a las conclusiones que pueden extraerse de un estudio. Así pues, problemas de gran tamaño suponen tensión ante todo sobre el lado del Procesamiento y Gestión de los Datos en la actualidad y no tanto sobre el del Análisis. Nos referimos, naturalmente, a esfuerzo físico o económico. El Análisis de Datos puede requerir mucho esfuerzo de tipo conceptual, en todo caso.

• El contexto del Procesamiento de los Datos. Bergman y Magnusson (1990) ofrecen la siguiente clasificación de fuentes de datos.

a) Datos provinientes de registros preestablecidos (Instituciones, empresas, hospitales, etc.).

b) Procedimientos de laboratorio y otros procedimientos que producen medidas directas (p.e. medir la longitud, el peso, los niveles de hormonas de una muestra de sangre).

c) Inventarios (p.e. un cuestionario acerca de condiciones y actitudes hacia el trabajo).

d) Tests (p.e. un test de inteligencia en el cual hay un número de preguntas normalizadas que deben ser contestadas como correctas o falsas).

e) Observaciones directas (p.e. de niños en situaciones de juego llevando a recuentos simples del número de veces que cierta conducta es mostrada).

f) Entrevistas (p.e. una entrevista psiquiátrica).

Cada una de estas situaciones implica unas exigencias más o menos diferentes desde el punto de vista del Proceso de Datos. Comentaremos a continuación algunos de las aspectos más relevantes para cada una de ellas.

Los datos provinientes de registros preestablecidos incorporan la dificultad muy a menudo de realizar traducciones entre los formatos utilizados originalmente por la institución o entidad que los produjo y aquellos que el investigador requiere para su trabajo. Esto puede suponer dos retos. En primer lugar, tenemos lo referido a la

trasferencia física de los datos, ya que, a menudo, las instituciones de cierto nivel desarrollan sus propios sistemas informáticos sin planear en absoluto su integración con otros diferentes. De este modo, los campos de las bases de datos a menudo no están bien delimitados, su tamaño puede ser incorrecto o, simplemente, no existe un mecanismo predeterminado para su exportación. Además, es posible que se utilicen protocolos de compresión que harían muy difícil acceder a la información utilizando una "puerta trasera". El otro problema sería la estructura lógica o conceptual. Puesto que los datos recogidos por la institución no fueron diseñados teniendo en cuenta las necesidades de una investigación particular, sino las suyas propias, a menudo es necesaria una tarea de codificación y recodificación que proporcione una nueva coherencia desde el punto de vista de nuestros intereses.

Los procedimientos de laboratorio y otros que producen medidas directas dan a lugar a complejidades con respecto a las transformaciones. Así, medidas de individuos que ofrecen resultados sucesivos de una manera casi continua (p.e. registros electrofisiológicos) pueden requerir de una gestión de datos relativamente compleja. Otras medidas directas producirán valores que no necesariamente siguen la distribución normal, así que el concepto de escalera de potencias (Emerson, 1991) puede resultar de gran importancia aquí.

Los cuestionarios suponen un reto importante para muchos aspectos del Proceso de Datos, así que no resulta extraño que se hayan convertido en el ejemplo prototípico de los libros referidos a él. En esta situación nos encontramos con problemas de codificación, de introducción de datos (ya que los cuestionarios pueden ser muchos y tener una gran casuística en cuanto a las preguntas), de calidad en los datos (a menudo los sujetos no desean contestar la verdad, o los entrevistadores cometen errores, o la introducción se realiza bajo presiones que producen efectos indeseados), se dan muchos casos de no - respuestas que hacen plantearse la necesidad de recoger los datos de nuevo o de realizar asignaciones, y, finalmente, es necesario hacer transformaciones y recodificaciones que combinen las preguntas relacionadas (p.e. escalas obtenidas a partir de preguntas relacionadas entre sí).

Los tests pueden ser vistos como versiones simplificadas de cuestionarios desde el punto de vista del Proceso de Datos. En ellos, todas las preguntas tienen el mismo formato (p.e. dos o cinco alternativas) y las transformaciones necesarias se limitan a la combinación de las subescalas que los componen.

Las observaciones directas presentan el problema de tener una estructura de datos inusual desde el punto de vista de los programas de gestión de datos. Sanderson (1994)

por ejemplo realiza una revisión de los diferentes programas que han sido utilizados para gestionarlos y propone un formato diferente que permitiría representarlos de manera adecuada sin realizar transformaciones. Sin embargo, en la práctica este tipo de datos sufre una gran cantidad de transformaciones del tipo de recodificaciones en busca de un nivel de análisis adecuado para ellos. Una transformación muy habitual es, por ejemplo, la obtención de matrices de transición entre estados secuencialmente situados a uno, dos o más pasos de cada estado. También, dado que estos estudios detectan dependencias secuenciales muy rígidas, y por tanto, poco interesantes, es necesario llevar a cabo recodificaciones que depuren las categorías elegidas para ser observadas hasta determinar cuáles son aptas para ser analizadas.

Finalmente, las entrevistas, aunque consideradas por Bergman y Magnusson como algo separado, desde el punto de vista del Procesamiento de Datos están muy cercanas al caso de los Inventarios. Un caso particular de gran interés es por otro lado, cuando el entrevistador es sustituido por un ordenador encargado de recoger las respuestas de los entrevistados (Saris, 1991).

## **La atención dedicada al Proceso de Datos**

Davidson (1996) hace notar que el número de textos disponibles que traten cuestiones básicas de gestión de datos es muy reducida comparado con los disponibles acerca de Estadística o Análisis de Datos.

Cabría preguntarse la razón por la que estos contenidos están siendo ignorados. Generalmente, los contenidos de Estadística o similares han sido enseñados en clases magistrales utilizando pequeños ejemplos que pueden ser seguidos utilizando una calculadora o incluso sin ella. Cuando los estudiantes han adquirido conocimientos básicos es posible acudir a un aula dotada de ordenadores en los que se pueden poner en práctica los conocimientos correspondientes. Estas clases prácticas implican varios elementos de dificultad. En primer lugar, los alumnos necesitan conocer los rudimentos del uso de un ordenador: usar un teclado, un ratón, manejar ventanas, gestionar mecanismos de almacenamiento, etc. Este problema cada vez es menos importante debido a que cada vez hay más oportunidades de adquirir estas habilidades y también a la progresiva simplificación de la forma en que se manejan los ordenadores. En segundo lugar, necesitan un conocimiento básico acerca del uso de un paquete estadístico, pero, de nuevo, esto es cada vez menos un inconveniente puesto que estos programas son actualmente bastante fáciles de usar y las habilidades requeridas no difieren de las necesarias para usar otro tipo de programas. En tercer lugar, necesitan conocer la forma

de crear un fichero de datos y depurarlo hasta estar listo para empezar a realizar análisis. Ahora bien, este último elemento resulta mucho más difícil de solventar por lo que generalmente es evitado usando alguna de estas dos estrategias: utilizar conjuntos de datos pequeños bien controlados o utilizar archivos previamente depurados y bien dispuestos que permitan comenzar la tarea de análisis desde el primer momento. Esto es comprensible puesto que no tendría sentido en una clase práctica llevar a cabo explicaciones que se desvíen mucho de lo explicado en teoría, por lo que es preferible hacer este salto e ir directamente a la cuestión. Como este tipo de contenidos tampoco parece encajar dentro de la asignatura de teoría, el resultado es que son evitados, dando lugar a una especie de salto en el aire que pone a los estudiantes en el lugar correcto sin conocimiento del proceso previo.

El apartado de contenidos de este texto es nuestro intento por llenar este hueco. En él se presenta material que cubriría los aspectos teóricos acerca de las tareas de recogida y gestión de datos, proporcionando asimismo un fundamento a una docencia de tipo práctico. El nivel aquí presentado es de tipo medio, superior al que un estudiante que deseara conocer sólo un mínimo suficiente requeriría, pero inferior al que alguien que se enfrente con problemas complejos puede necesitar. Esta situación es inevitable puesto que resulta muy difícil cubrir todas las necesidades posibles con un único texto y nuestro objetivo ha sido el de ofrecer un esquema desarrollado mínimamente que, en caso de ser considerado como acertado, pueda formar la base de otros esfuerzos futuros más ajustados a las distintas necesidades.

Una cuestión que puede producir cierta extrañeza es haber dado tanta importancia a la teoría. A menudo se percibe que estos contenidos están vinculados con aspectos de tipo práctico y que el único lugar adecuado para ser impartidos es un aula con ordenadores. No obstante, a partir de nuestra experiencia las clases prácticas tienen un carácter tan concreto que, sin una contextualización adecuada, pueden resultar muy poco útiles. Puesto que las clases prácticas deben estar centradas en los detalles concretos que imponen el equipamiento disponible, muy a menudo es necesario dar importancia a cuestiones que son completamente accesorias y que pueden cambiar en cualquier momento debido a, por ejemplo, la renovación del software o el hardware disponible. Así, sin una conceptualización que permita generalizar lo aprendido a otros contextos, los conocimientos adquiridos en un aula de informática tendrían un alcance muy limitado, ligado a las condiciones en las que fueron aprendidos.

Un aspecto relacionado es la no vinculación a un determinado sistema o programa de gestión de datos. Aunque en las clases prácticas es necesario centrarse en un único sistema, nuestra intención ha sido plantear el contenido de este texto sin hacer referencia a

cuál es el elegido en concreto. Esto presenta dos ventajas. En primer lugar, cambios o renovaciones del software no supondrán que el material aquí presentado pase a ser obsoleto, y, en segundo lugar, nos ha permitido plantear los temas con total libertad, sin las constricciones impuestas por la determinada forma de hacer las cosas impuesta por el programa informático elegido. Obviamente, el material de las prácticas entregado finalmente a los alumnos debería tener un grado de concreción mayor, pero nuestra experiencia durante los últimos años nos ha resignado a tener que rehacer parte de él continuamente, para adaptarnos a los cambios que ha sufrido y sufrirá la tecnología de los ordenadores personales.







# ***Definición y Estructuras de Datos***

## **1.1. Introducción**

Por proceso de datos entendemos los pasos que van desde el momento en que se recogen unos datos hasta que éstos están listos para ser analizados estadísticamente, o al menos, poder ser utilizados para responder a preguntas que interesen a los que los tengan a su disposición. Obviamente, este proceso será mejor cuanto más eficiente y libre de errores sea. Cuanto más grande es el tamaño del conjunto de datos más problemas pueden surgir y las consecuencias en tiempo, coste o pérdida de calidad en los resultados pueden ser más graves.

Un supuesto prácticamente necesario para dar sentido a la mayoría de este texto es que los datos van a ser manipulados buena parte del tiempo por un computador. Aunque podríamos plantearnos un proceso de datos sin ordenador esto sería un anacronismo poco concebible en los tiempos actuales.

En el gráfico 1.1 se representan los pasos del proceso de datos que hemos distinguido en este texto.

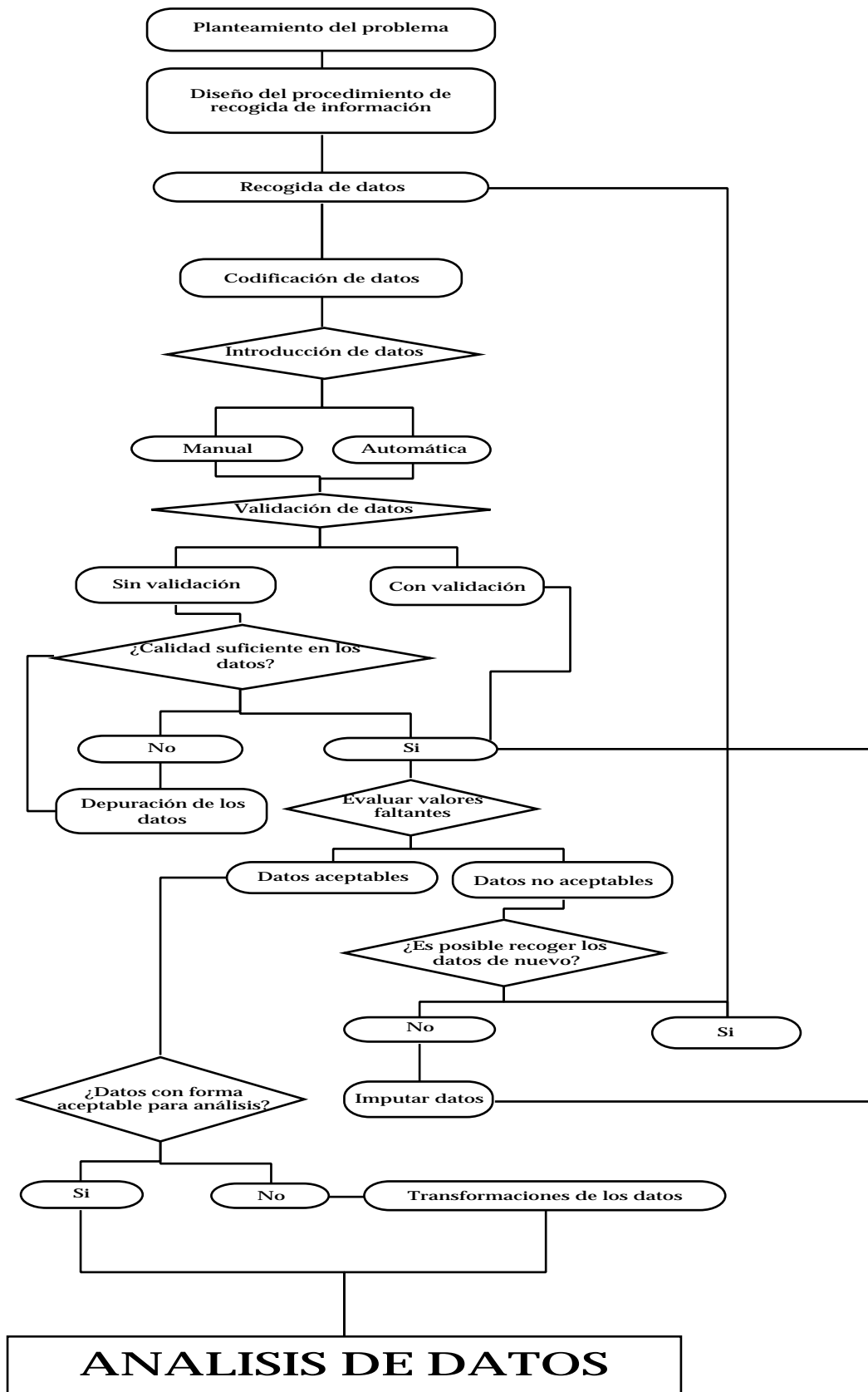


Figura 1.1. El esquema del Proceso de Datos

En el esquema anterior tendríamos que el primer paso en cuanto a la parte de procesamiento de datos sería el planteamiento del problema y el diseño del procedimiento de recogida de la información. Estas dos tareas no han sido consideradas en este texto como propias del procesamiento de los datos ya que implican conocimientos metodológicos que seguramente están mejor cubiertos en otras asignaturas. A continuación se pasaría a la fase de recogida de los datos. Esta fase tampoco se ha explicado con detalle.

El capítulo 2 de estos contenidos se centra en la codificación de datos de cuestionarios. La razón para elegir este caso concreto como ejemplo didáctico se debe a que son datos altamente estructurados y de cierta complejidad, así como bastante comunes en muchos trabajos prácticos. Así, los principios generales acerca de codificación de datos podrían ser expuestos a partir de él.

El siguiente paso es el de la introducción de datos. El capítulo 3 presenta una exposición de diversos métodos para llevar a cabo esta tarea que permitirían comprender el abanico de posibilidades existente para aquellos que llevaran a cabo tareas de procesamiento de datos de manera continua o profesional. De este modo, una distinción importante a realizar desde el principio es si la introducción de datos va a hacerse utilizando métodos automáticos o manuales y posteriormente si se va a utilizar un método de validación que depure los errores detectables. En el capítulo 4 se exponen los principios básicos de una tarea que, aunque no mostrada en ninguna parte del gráfico podría ser necesario realizar en este momento (aunque en realidad seguramente será necesitada en tantos otros momentos que, para no hacer excesivamente complejo el gráfico no ha sido señalada): importar/exportar los datos.

En la práctica, una fase de depuración de los datos es siempre aconsejable, así pues el camino a seguir debería transitar al menos una vez por este paso. El capítulo 5 muestra técnicas de comprobación de la calidad de los datos.

Posteriormente resulta aconsejable evaluar los datos faltantes o ausentes para examinar hasta qué punto representan una distorsión de la matriz de datos que haría los análisis difíciles de interpretar y/o aceptar. Cuando los datos no sean aceptables sería conveniente plantearse una nueva recogida de los mismos o, en su defecto, una asignación de datos que posteriormente debería ser reevaluada de nuevo. El capítulo 6 muestra una introducción al tema de los datos faltantes, los problemas que se pueden derivar de ellos y algunas técnicas para la asignación de éstos.

Una vez los datos son considerados aceptables pasaríamos a realizar una serie de transformaciones de éstos en caso de considerar que no están listos para su análisis. Este tipo de transformaciones resultan necesarias por ejemplo cuando el nivel de recogida de los datos es diferente del nivel de análisis. También, ciertas operaciones requeridas por el instrumento de recogida de datos pueden ser relegadas al ordenador, con el consiguiente ahorro de tiempo y esfuerzo por parte del operador humano y la mejora en precisión y velocidad que caracteriza la ejecución de los computadores.

Finalmente, puesto que muchas de esas tareas pueden ser automatizadas y ejecutadas de una manera conjunta mediante un lenguaje de programación dedicaremos el último capítulo a revisar los conceptos básicos en relación con esta cuestión.

Volviendo de nuevo al primer capítulo, indicar que en él nos centraremos primordialmente en cuestiones introductorias acerca de los datos. Discutiremos en primer lugar una definición sencilla de **datos** para posteriormente pasar a las fuentes de los datos. En último lugar se encuentra el apartado más complejo de ésta sección, dedicado a las **estructuras** de los datos, ya que este conocimiento es importante para llevar a cabo las tareas de procesamiento de datos. Estas estructuras de datos pueden ser entendidas también desde el punto de vista del computador y la forma que éste ofrece para manejarlas, siendo tratado este tema en un último apartado.

## 1.2. Definición de datos

A lo largo de este texto la palabra datos aparecerá de una manera continua. Resulta pues conveniente introducir algunos comentarios acerca de su significado. Puesto que nuestro objetivo está más centrado en el procesamiento que en cuestiones epistemológicas, no entraremos a considerar el concepto de datos con mucha profundidad. Coombs (1964) suele ser nombrado como una referencia clásica para una discusión rigurosa del término.

Velleman (1995) comenta que el término **datos** no suele ser definido en los libros de Estadística y que son los manuales más orientados al software los que tienen más interés en aclarar su sentido. Para él **datos** son "valores midiendo o dando a conocer información acerca de cada individuo en un grupo junto con detalles acerca de como los valores se relacionan entre sí y acerca de qué dan a conocer información". Esta definición vemos que enfatiza el hecho de proporcionar información y el que los datos deben ser *acerca de algo* y no solamente una mera colección de números o nombres sin un contenido.

La definición anterior es demasiado simplista en nuestra opinión ya que se centra excesivamente en lo que se conocen como datos rectangulares tal y como son explicados en un apartado posterior. Ello deja fuera por ejemplo los datos acerca de cómo una entidad se relaciona con otra entidad (datos de similitud).

Federer (1982) plantea la necesidad de distinguir entre números, adjetivos u otras formas de descripción y datos. Los números y los adjetivos pueden estar disponibles por sí mismos o como datos, pero la existencia de los primeros no implica tener un conjunto de datos. No obstante, un conjunto de datos siempre implica la existencia de alguna forma descriptiva tal y como números, adjetivos, frases, dibujos, gráficos, etc. Por ejemplo, el conjunto de números {3, 1, 0, 4, 9, 6} no transmite en sí mismo información acerca de un fenómeno pero cuando éstos se ponen en relación con un determinado fenómeno (por ejemplo, número de gusanos en una manzana), podemos considerarlos como datos.

Para Young (1987) los datos son el resultado de un proceso de clasificación. Es decir, los datos son siempre categóricos. Para él, el supuesto básico es que siempre es posible para dos datos decidir si son equivalentes o no. Asimismo, también se asume que los datos son recogidos en una situación empírica con unas características conocidas. Las relaciones entre observaciones forman para él parte de las características de medición de los datos, algo que puede ser tratado aparte y que determina si las relaciones entre las diferentes observaciones pueden ser consideradas de un cierto nivel u otro (así como si el proceso de medida puede ser considerado discreto o continuo). Otra característica importante de los datos es su forma, lo cual hace referencia a las condiciones, niveles, replicaciones, etc., en que los datos se encuentran dispuestos.

Así pues, podemos resumir las definiciones anteriores en el sentido de señalar que **datos** son formas de descripción que tienen al menos un carácter categórico (es decir, es posible al menos indicar si dos datos son iguales o diferentes), dispuestos en una estructura que hace referencia al diseño con el que fueron recogidos, que transmiten información de interés y que pueden tener unas características de medición dadas.

En un apartado posterior se tratará el tema de las estructuras de datos, cuestión que es fundamental a la hora de disponerlos para poder ser analizados. Los temas de diseño de recogida de datos y de características de medición de los datos no serán tratados en este texto ya que nos parece que corresponden mejor a textos con una orientación diferente (p.e. Diseño Experimental o Psicometría).

### 1.3. Fuentes de datos

La procedencia de unos datos impone una serie de condicionantes sobre la forma que éstos tendrán, las dificultades que tendremos para gestionarlos y posiblemente el tipo de conclusiones que se puedan extraer.

Bourque y Clarke (1992) señalan las siguientes fuentes:

- **Cuestionarios.** Los cuestionarios son conjuntos de preguntas acerca de, por ejemplo, actitudes, opiniones o conducta presente, pasada o futura. Desde el punto de vista del procesamiento de datos (Swift, 1996) este es un método que puede decirse que produce datos claramente estructurados siendo su tratamiento un proceso bien definido. No obstante, debido a la gran casuística de preguntas que se suele dar son relativamente más complejos que un caso particular de aquellos, los tests. En los ítems, las preguntas suelen tener un aspecto similar con, por ejemplo, un número similar de categorías de respuestas, formuladas en la misma dirección, o, al menos, fácilmente reversibles para conseguir este objetivo.

- **Observaciones.** La observación puede ser un método que incluye muy diferentes formas de recoger datos en la práctica. En un extremo, la observación puede referirse a un aspecto muy concreto de la variable considerada obtenida en una situación altamente organizada. Por ejemplo, un investigador puede observar una medida del tiempo de reacción de un sujeto tras la presentación de un estímulo (generalmente de un modo indirecto a partir de su grabación por medio de algún aparato y posterior comprobación). En otra forma de observación, el experimentador registra los sucesos a medida que se van produciendo siendo en cierto modo él mismo el instrumento de medición. En el primero de los casos, el procesamiento de los datos no suele ser visto como un problema (lo cual no quiere decir que este texto no tenga nada que ofrecer a los que realizan este tipo de investigaciones), puesto que el instrumento de medición generalmente proporciona un registro fácil de usar y, al estar limitado a unas pocas variables, sencillo de disponer adecuadamente. En el segundo de los casos, la preparación, organización, estructuración y codificación de datos son fases tan importantes de la investigación que suele ocupar buena parte de los esfuerzos de aquellos que las llevan a cabo.

- **Registros.** Una forma de investigación es la que se realiza a partir de las trazas o huellas producidas por la actividad del sujeto u objeto investigado. En el caso de las Ciencias Sociales y sobre todo en las últimas décadas, multitud de instituciones recogen de una manera habitual información acerca de sus usuarios, clientes o miembros, la cual puede ser utilizada para responder preguntas sustantivas. La mayor ventaja de este tipo de

datos es su disponibilidad. A menudo, cuando es posible acceder a ella, el investigador se encuentra con una fuente de datos enorme y de gran riqueza. Su mayor desventaja es que la recogida de datos no fue planteada teniendo en cuenta sus necesidades, sino la de las instituciones que la llevaron a cabo. Esto significa que, por ejemplo, determinadas preguntas no podrán ser respondidas con total rigor o, quizás, ni siquiera investigadas. Además, desde el punto de vista del procesamiento de los datos nos encontraremos con registros a menudo confusos, categorías mal definidas, información separada en varios archivos, etc.

Como una clasificación alternativa podemos mencionar la mostrada por Bergman y Magnusson (1990). Estos distinguen entre:

- Recogida de datos de registros (p.e. datos escolares, delitos, etc.)
- Procedimientos de laboratorio y otros produciendo medidas directas (p.e. la medida de la altura y el peso, niveles hormonales de muestras de sangre, etc.).
- Inventarios (p.e. cuestionarios acerca de las condiciones de trabajo).
- Tests (p.e. un test de inteligencia en el que hay una serie de preguntas estandarizadas que pueden contestarse correcta o incorrectamente).
- Observaciones directas (p.e. de niños en situaciones de juego que producen recuentos simples del número de veces que se produce cierta conducta u otras apreciaciones más globales de ésta).
- Entrevistas (una entrevista psiquiátrica que produce una valoración o una entrevista estandarizada acerca de condiciones de trabajo).

Esta clasificación es muy similar a la de Bourque y Clarke. Sólo que separa de una manera más clara los datos provenientes de las diversas fuentes. No obstante, los comentarios anteriores podrían también ser aplicables en este caso.

## **1.4 Estructuras de datos**

Young (1987) señala que los datos tienen una serie de características que pueden ser divididas a nuestra conveniencia en dos grupos. Las referidas a la forma de los datos y las referidas a las características de medición. En nuestro caso, sólo atenderemos al primer grupo de características y no a las referidas a las características de medición (las cuales encajan más dentro de un texto referido a Psicometría por ejemplo). Determinar las

posibles formas de los datos que podrían ser recogidos es de interés desde el punto de vista del procesamiento y gestión de datos porque nos delimita las estructuras que será razonable obtener en la práctica. De este modo, aunque la teoría acerca de bases de datos permita plantear estructuras de muchos tipos, nosotros sólo utilizaremos una parte limitada de ellas.

ESCALAS

OLORES


Figura 1.2. Ejemplo de datos de dos vías.

Carroll y Arabie (1980) y Young y Hammer (1987) indican características muy similares para la forma de los datos. Lo siguiente es una mezcla de ambos pero más centrada en el segundo. Son las siguientes:

1. *Número de vías.* Esto se refiere al número de condiciones experimentales o factores en los datos. Otros términos que son utilizados con significado similar (aunque Young prefiere reservarlos para otros aspectos de la teoría de datos) podrían ser variables independientes, componentes, dimensiones, facetas, modos, etc.

P A E

Sujetos


Figura 1.3. Ejemplo de datos de dos vías siendo una de ellas sujetos



Las observaciones o datos serían elementos del producto cartesiano de todos los niveles de las condiciones experimentales.

Algunos ejemplos podrían ser: si uno tiene observaciones acerca de la valoración de cuatro olores en seis escalas (atracción, dulzura, masculinidad, etc.) estaríamos ante datos de dos vías. Este diseño sería el producto cartesiano  $< 4 \times 6 >$  y estaría representado en la figura 1.2. Las observaciones estarían en las celdas.

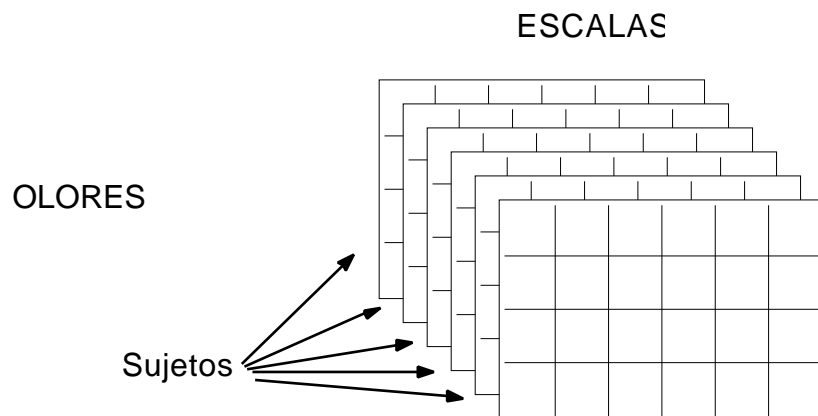


Figura 1.4. Datos de tres vías

El ejemplo anterior es un poco inusual en ciencias sociales, ya que muy a menudo una de las vías serán sujetos. Por ejemplo, supongamos que tenemos como una de las condiciones medidas de peso (P), altura (A), y edad (E) de seis sujetos. Los datos se organizarían como en la figura 1.3 y serían de nuevo datos de dos vías.

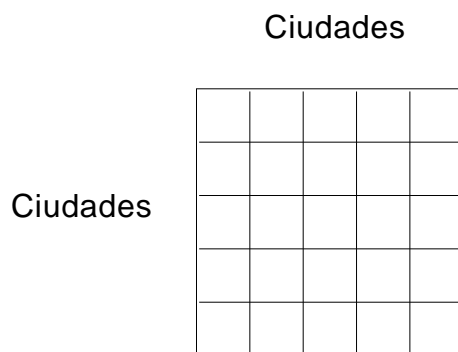


Figura 1.5. Distancias entre ciudades

Combinando ambas situaciones podríamos tener datos de tres vías. Si preguntamos a un grupo de sujetos que califiquen una serie de olores en una serie de escalas tendríamos datos con la forma de la figura 1.4.

2. Número de modos. Una situación que puede surgir en nuestros datos es que dos vías sean iguales. Por ejemplo, supongamos que tenemos las distancias que hay entre cinco ciudades. La figura 1.5 podría ser una representación gráfica de esta situación:

Parece claro que es conveniente distinguir este caso de aquel en que las vías son diferentes entre sí. Para ello se utilizará el concepto de modo. El número de modos de unos datos hace referencia al número de vías diferentes en unos datos. Así, los ejemplos de las figuras 1.2 y 1.3 corresponderían a datos con dos vías y dos modos y el de la figura 1.5 sería de dos vías y un modo. Otro ejemplo, si preguntamos a una serie de sujetos acerca de qué distancia creen que hay entre cinco ciudades en un país. Los datos corresponderían a tres vías <ciudades x ciudades x sujetos> y dos modos (ciudades y sujetos) y podrían representarse como se muestra en la figura 1.6. Arce (1994) expone procedimientos de recogida de datos que dan lugar a estructuras de este tipo.

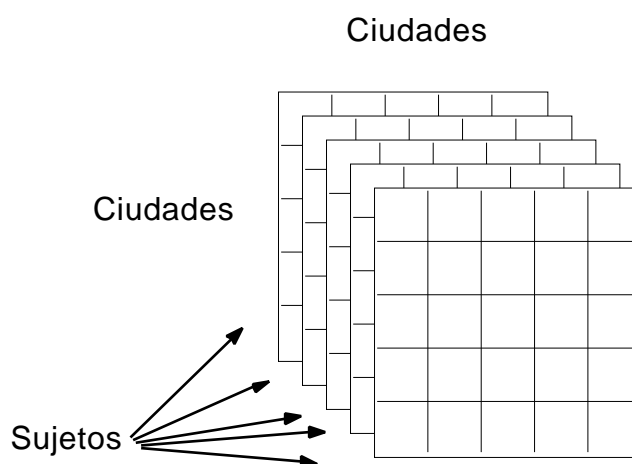


Figura 1.6. Datos de tres vías y dos modos

Llegado a este punto conviene introducir una terminología utilizada muy habitualmente para describir la forma de unos datos. En primer lugar, tendríamos que datos *rectangulares* son datos de dos vías y dos modos (aunque por casualidad suceda que las dos vías tienen el mismo número de niveles seguiremos utilizando el nombre de datos rectangulares) y datos *cuadrados* son datos de dos vías y un modo. Los datos denominados *multivariados* corresponderían a datos rectangulares en que muy a menudo

se asume que una de las vías corresponde a sujetos. A menudo los datos de tres vías son etiquetados como cuadrados o rectangulares. Ello haría referencia respectivamente a si el número de modos es menor que el número de vías o inferior. El ejemplo de la figura 1.6 sería por tanto datos cuadrados de tres vías.

3. Simetría. Unos datos de dos vías y un modo (la característica de simetría sólo se aplica a datos cuadrados) no es simétrica si la parte inferior izquierda no es igual respecto de la parte superior derecha. Por ejemplo, supongamos que en la matriz de distancias entre ciudades introducimos tiempo necesario para ir de una a otra por carretera (y tenemos en cuenta que en algunos casos será más rápido ir desde la ciudad X a la ciudad Y que volver desde la ciudad Y hasta la ciudad X debido a circunstancias propias de cada una de ellas). En este caso la matriz será asimétrica.

	PENACAP	CACHETE	HORASTV	LEYARMAS
PENACAP	1.000	-.148	.088	-.108
CACHETE	-.148	1.000	-.028	.055
HORASTV	.088	-.028	1.000	-.065
LEYARMAS	-.108	.055	-.065	1.000

*Figura 1.7. Ejemplo de matriz simétrica*

En la figura 1.7 tenemos una matriz simétrica. En ella se muestra la matriz de correlaciones de Pearson entre las variables PENACAP (estar de acuerdo con la pena de muerte), CACHETE (estar de acuerdo en dar algún cachete a los hijos como parte de su educación), HORASTV (número de horas que ve la televisión) y LEYARMAS (estar de acuerdo con la ley que permite a los ciudadanos llevar armas) para una muestra de ciudadanos de Estados Unidos.

	PENACAP	CACHETE	HORASTV	LEYARMAS
PENACAP	.167	-.054	.086	-.017
CACHETE	-.054	.797	-.061	.018
HORASTV	.086	-.061	5.827	-.059
LEYARMAS	-.017	.018	-.059	.141

*Figura 1.8. Ejemplo de matriz simétrica con diagonal*

Puesto que muchos programas de análisis estadístico sólo utilizan la mitad inferior o la mitad superior de este tipo de datos es posible hacer una distinción entre datos

triangulares inferiores o superiores, siendo la diagonal el punto de separación entre ambos triángulos.

También, la diagonal puede o no ser digna de considerar. Por ejemplo, en el caso de la matriz de la figura 1.6 anterior, la diagonal no es importante puesto que por definición es siempre igual a 1. Por el contrario, si estuviéramos trabajando con la matriz de varianzas-covarianzas de las mismas variables tendríamos la figura 1.8. En ella, la diagonal corresponde a las varianzas de las variables y las otras casillas a las covarianzas.

4. Datos completos. Un conjunto de datos puede ser completo o no. Un diseño es

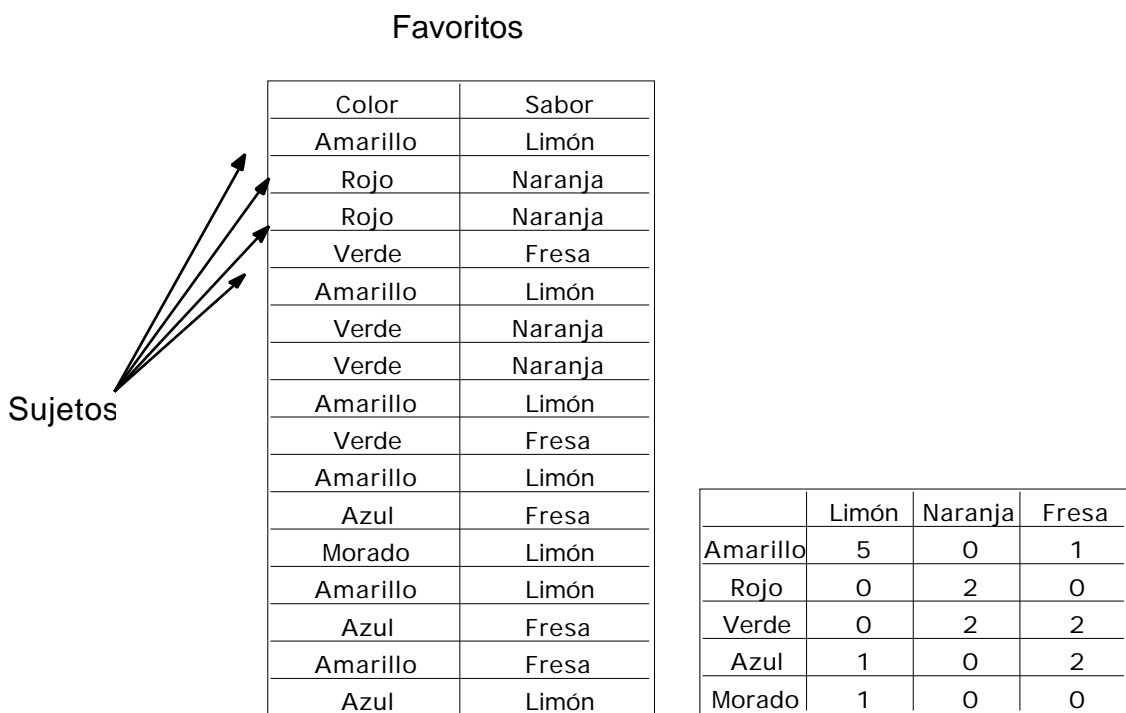


Figura 1.9. Datos de una tabla de contingencia antes y después de agregar

completo cuando cada celda del diseño contiene una observación. Así, cuando hay datos faltantes el diseño es incompleto. Los diseños incompletos pueden serlo tanto si siguen un patrón predeterminado (planeado) como si no. Los patrones incompletos planeados intentan garantizar la extracción de conclusiones a pesar de la falta de información. El término datos faltantes suele reservarse a cuando el patrón incompleto no ha sido planeado, aunque a veces los términos se entremezclan (Graham et al., 1996).

5. Datos agregados. Este tema no está expuesto por Young (1987), siendo, quizás, innecesario para caracterizar la forma de los datos, aunque útil para entender ciertas situaciones. Tenemos datos agregados cuando podemos concebir que la matriz que

contemplamos puede ser entendida como una transformación de datos dispuestos a un nivel más inferior o detallado, tal que la información a ese nivel individual se ha perdido (estos datos se denominan microdatos). Por ejemplo, la matriz de correlaciones de la figura 1.7 es el resultado de transformar una matriz de dos vías y dos modos (sujetos x variables) en una matriz de dos vías y un modo (variables x variables). Este caso correspondería a la pérdida de un modo, la información de las variables referida a cada sujeto. Por otro lado, la matriz de correlaciones nos permite interpretar más claramente ciertos aspectos de nuestros datos que de otro modo resultaría difícil de ver.

Una situación muy común de datos agregados son las tablas de frecuencias cruzadas. En ellas podemos pensar que los datos originales son los situados en la parte izquierda de la figura 1.9 y los datos agregados los situados en la parte derecha. En este caso no hemos perdido ningún modo, ni vía. Sin embargo, puesto que el número de observaciones original era mucho mayor podemos percibir igualmente que se ha perdido cierta cantidad de información (aunque se ha ganado otra). La agregación de datos es un tema de gran importancia estadística con graves consecuencias en caso de no tratarse bien.

## **1.5. Casos especiales**

En este apartado se discuten algunos tipos de datos que no encajan con facilidad en las estructuras propuestas en la sección anterior. En concreto, tenemos datos textuales (a veces llamados simplemente datos no estructurados) y los datos secuenciales.

### **1.5.1. Datos textuales.**

En la definición de datos hemos comentado que éstos deben tener estructura para ser considerados como tales. Unos datos que se encuentran en la frontera son los datos de texto. Estos datos son muy habituales dentro de las estrategias de investigación cualitativas (Boulton y Hammersley, 1996; Miles y Huberman, 1994; Weitzman y Miles, 1995) aunque también pueden aparecer en estrategias de investigación cuantitativas en las que se utiliza información del tipo preguntas cortas.

En este tipo de información resulta complicado distinguir entre la realidad que estamos estudiando y los datos que hemos extraído de ella. Ello se debe a que muy a menudo los textos que disponemos, y suponen la materia prima del estudio, deben pasar por una serie de codificaciones y recodificaciones que los dispongan de una manera apropiada para ser analizados. De este modo, los textos no son considerados ellos mismos como datos, sino un material preliminar del cual extraer esos datos. No obstante,

resulta difícil hablar en términos generales, debido a la gran diversidad de orientaciones teóricas existentes (Miles y Huberman, 1994).

En este trabajo se asumirá que los datos estudiados pueden ser dispuestos en una estructura como la de la figura 1.10.

En esa representación cada sujeto emite una contestación acerca de uno o varios temas. Nuestro supuesto es que cada una de esas contestaciones es relativamente corta y además, a menudo, no hay muchos temas por sujeto.

Otras estructuras de datos de texto más complejas no serán tratadas en este trabajo. La estructura anterior corresponde a un caso muy sencillo el cual tiene bastante importancia en la elaboración de encuestas ya que corresponde al de utilización de preguntas abiertas que han de ser codificadas para su análisis. Este tipo de preguntas serán tratadas en secciones siguientes.



Figura 1.10. Representación de datos textuales

### 1.5.2. Datos de categorías secuenciales.

La estructura de datos secuenciales presenta algunas peculiaridades interesantes. Utilizando la nomenclatura habitual diremos que estos datos suelen tener habitualmente dos, o más, tres vías. En el primer caso tendríamos sucesos y variables, y en el segundo incorporaríamos sujetos o unidades observadas. En la figura 1.11 tendríamos una representación del caso con tres vías, siendo el de dos vías simplemente uno de los rectángulos para un sujeto. No obstante, surgen complejidades que comentaremos más adelante.

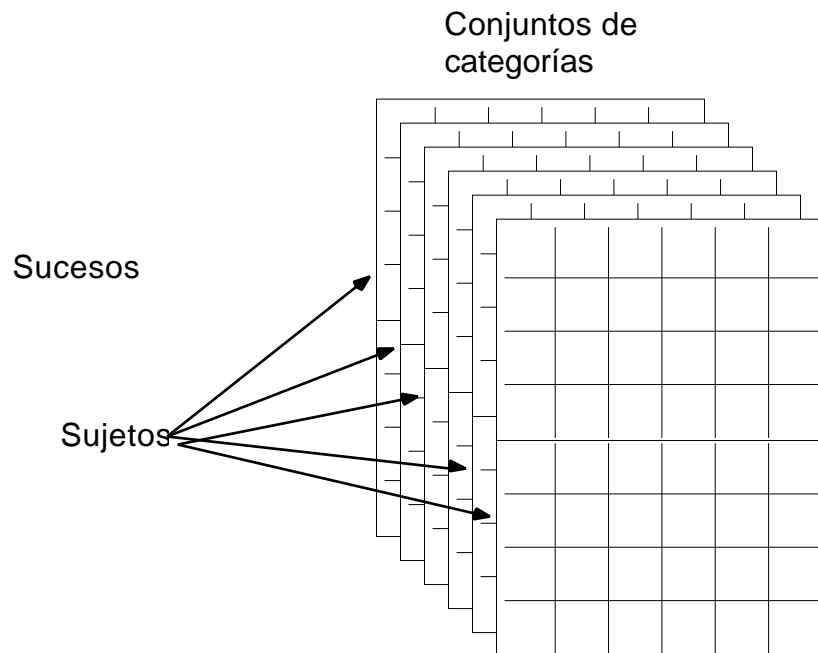


Figura 1.11: Datos secuenciales para varios sujetos

Un ejemplo ayudará a concretar mejor la cuestión. Supongamos que tenemos dos niños en los que observamos variables correspondientes a los conjuntos de categorías JUEGO <juega, no juega> y COMPAÑÍA <sólo, acompañado>. Nuestro registro observacional para el primer sujeto podría tomarse durante el tiempo de recreo en la escuela y produciría para el primer niño algo así como JUEGO: j n j n j n j n ... etc. y COMPAÑÍA : s a s a s a ...etc. Cada una de las letras minúsculas corresponde a una de las categorías de las variables consideradas. Estas secuencias se dispondrían usando la figura 1.11 a lo largo del lado etiquetado como sucesos. Como tenemos dos sujetos tendríamos dos rectángulos como los de la figura 1.11.

Las complejidades que surgen para esos datos son las siguientes:

- Cada una de las variables tendría una longitud diferente para cada sujeto. Supongamos que un niño pasa por diferentes momentos de juego y no juego pero siempre está sólo. Tendríamos entonces en la columna JUEGO una cadena de sucesos y en la de compañía solamente uno.
- Las longitudes de columnas pueden ser diferentes para diferentes sujetos. Supongamos que el otro niño pasa por periodos de compañía y de soledad. En la variable

de de COMPAÑÍA tendría una cadena de sucesos mientras que el primero sólo tendría un suceso.

speechcode		state.var	
		1	00:00:00:00 00:00:00:00 (M,L,5)
		2	00:00:00:00 00:00:00:00 (I,A,5)
		3	00:00:00:00 00:00:00:00 (O,A,1)
		4	00:00:00:00 00:00:00:00 (H,A,5)
1	00:00:01:00 00:00:06:00 GOAL(+,+,+)		
2	00:00:07:00 00:00:09:00 COMMAND(Sam,Tom,+1)		
3	00:00:10:00 00:00:11:00 ACKNOWL(Tom,Sam,-1)		
4	00:00:11:00 00:00:13:00 ANNOUNCE(Sam,0)	5	00:00:11:00 00:00:12:00 (H,A,9)
5	00:00:13:00 00:00:21:00 ANNOUNCE(Tom,0)	6	00:00:19:00 00:00:21:00 (O,A,8)
6	00:00:21:00 00:00:24:00 OBSERVE(Sam,0)	7	00:00:21:00 00:00:23:00 (I,A,8)
7	00:00:24:00 00:00:26:00 OBSERVE(Sam,0)		

Figura 1.12. Representación de datos secuenciales con duración

• En muchos estudios se registra el tiempo que los sujetos están en la categoría correspondiente. En nuestro ejemplo, tendríamos por ejemplo JUEGO: j(0:20) n(20:23) j(23:45) n(45:46)...etc. En este caso los números entre paréntesis indicarían el tiempo en segundos en cada categoría. Esta codificación es interesante porque a partir de ella ahora podemos plantearnos por ejemplo en qué momentos el sujeto estaba jugando y acompañado simultáneamente (antes no podíamos). En este caso, la representación de la figura 1.11 podría ser válida aunque desde un punto de vista práctico quizás plantea ciertas dificultades de inspección visual. MacShapa (Sanderson y Fisher, 1994; Sanderson, et al., 1994) ofrece la representación de la figura 1.12. En ella, cada celda representa una categoría de observación dentro de una variable. El tiempo forma parte de la información dentro de cada observación y, para permitir la inspección visual, las celdas



se han hecho más grandes o más pequeñas proporcionalmente a su duración. También, cuando no hay solapamiento entre variables se han dejado huecos para representarlo.

Los datos secuenciales todavía presentan complejidades mayores. Un texto que plantea con más rigor los distintos tipos de datos secuenciales y que ofrece un esquema para su codificación es Bakeman y Quera (1995). Una revisión de software con consideraciones acerca de la forma de la información secuencial se puede encontrar en Sanderson (1994).

## **1.6. Estructuras de datos y computadores**

La exposición anterior muestra una clasificación de estructuras de datos abstracta, sin tener en cuenta cómo los paquetes estadísticos o los programas de gestión de datos pueden representar esas estructuras.

Las posibilidades de representación de los programas informáticos pueden ser clasificadas en externas e internas. Las externas hacen referencia a representaciones diseñadas para ser mostradas a los usuarios y están dominadas en la actualidad por el modelo de tabla rectangular (aunque algunos programas implementan otros). Las internas son mucho más flexibles y dentro de ellas describiremos la estructura relacional, la cual suele subyacer a un tipo de programas denominados de bases de datos, y las que se encuentran en lenguajes de programación.

### **1.6.1. Estructuras de datos con tablas rectangulares.**

Los datos en paquetes estadísticos generalmente se organizan en tablas rectangulares tal y como la mostrada en la tabla 1.13.

En el caso más simple tenemos datos de dos vías con una de ellas perteneciendo a sujetos, los cuales corresponden a las filas. El nombre de las variables suele aparecer en la parte superior de la tabla de datos. El cruce entre una variable y un sujeto corresponde con el valor de esa variable en ese sujeto. Este tipo de datos recibe el nombre de datos rectangulares o multivariantes y son los más habitualmente manejados en Ciencias Sociales.

Un segundo caso habitualmente tratado por los paquetes estadísticos es el de datos de dos vías y un modo. Estos datos suelen corresponder a datos de similitudes o distancias, correlaciones o covarianzas. Por ejemplo SPSS incluye las siguientes palabras claves para su comando *MATRIX DATA* (datos de matriz).

<i>Sujetos</i>	<i>H<sup>a</sup></i>	<i>Mat.</i>	<i>Gen</i>	<i>D-48</i>
A. García	2.5	2	M	22
B. Martín	2.5	3	H	25
L. Perez	2	1	H	22
R. Sánchez	2.5	1	M	22
S. Perez	3	1	H	31
G. Molina	2.5	4	H	21
D. Monzo	1.5	1	M	28
R. Santos	2.5	1.5	H	22
M. Pozo	3.5	1	H	22

Tabla 1.13. Tabla de sujetos por variables.

- *LOWER* (inferior): La matriz sólo tiene valores en la parte inferior. Esto correspondería a datos simétricos.
- *UPPER* (superior): La matriz sólo tiene valores en la parte superior. Esto correspondería igualmente a datos simétricos.
- *FULL* (completa): La matriz está completa. Esto puede corresponder a un matriz simétrica con información redundante o a una matriz asimétrica.
- *DIAGONAL/NODIAGONAL*: La matriz incluye o no la diagonal. La diagonal puede ser omitida cuando contiene información conocida, como, por ejemplo, cuando se trata de matrices de correlaciones.

También, ciertos paquetes estadísticos admitirán datos de matriz organizados como una columna o una fila de datos. Por ejemplo, la matriz de correlaciones de la figura 1.14 puede ser aceptada por un paquete estadístico tal y como se muestra en la columnas situadas a su izquierda.

Un tercer tipo de datos que aceptan algunos paquetes estadísticos son los datos de tres vías y dos modos. Estos datos pueden corresponder a matrices de similitudes para diferentes entidades o sujetos. En este caso, la representación habitual es situar una matriz de similitudes encima de la otra. ViSta (Young, 1996), por ejemplo, es un programa que incorpora facilidades para el manejo de este tipo de datos y que muestra el aspecto de la figura 1.15. En el ejemplo se muestran los juicios de similitud entre una serie de bebidas con sabor a cola para 10 sujetos. Sólo los tres primeros y parte del cuarto son mostrados. Los nombres de los sujetos mostrados son respectivamente Y1, N2, N3 e Y4. ViSta incorpora facilidades para la introducción de estos datos que no suelen ser contempladas en otros programas, tal y como añadir una matriz de datos completa (es decir, un nuevo sujeto) a la tabla.

1
0,25
1
0,56
0,21
1
0,23
0,58
0,7
1

0,25
0,56
0,21
0,23
0,58
0,7

1			
0,25	1		
0,56	0,21	1	
0,23	0,58	0,7	1

Figura 1.14. Matriz de correlaciones en forma de lista.

Los datos con estructura relacional suelen ser citados como un ejemplo de las limitaciones que ofrecen las representaciones en forma de tabla para cubrir ciertas situaciones (Davidson, 1996, ICPSR, 1997). Un ejemplo de estructura relacional (posteriormente mostraremos una explicación más detallada de este caso) podría ser el siguiente: En una encuesta se solicita la edad del cabeza de familia de cada hogar y la de los hijos. Digamos que el número de hijos por familia resulta un dato difícil de anticipar por lo que se guarda espacio para un total de 10 hijos (un margen muy generoso para la mayoría de las familias). El resultado es una tabla de datos multivariados de la cual mostramos una parte en la tabla 1.16. Esta tabla presenta dos inconvenientes. En primer lugar, produce la apariencia de que falta mucha información, la cual ha tenido que ser llenado con un símbolo cualquiera. Esto es un problema para la realización de análisis posteriores. En segundo lugar, hay un enorme desperdicio de espacio que redundará en una mayor necesidad de sistemas de almacenamiento y en una mayor incomodidad en cuanto al manejo de la información.

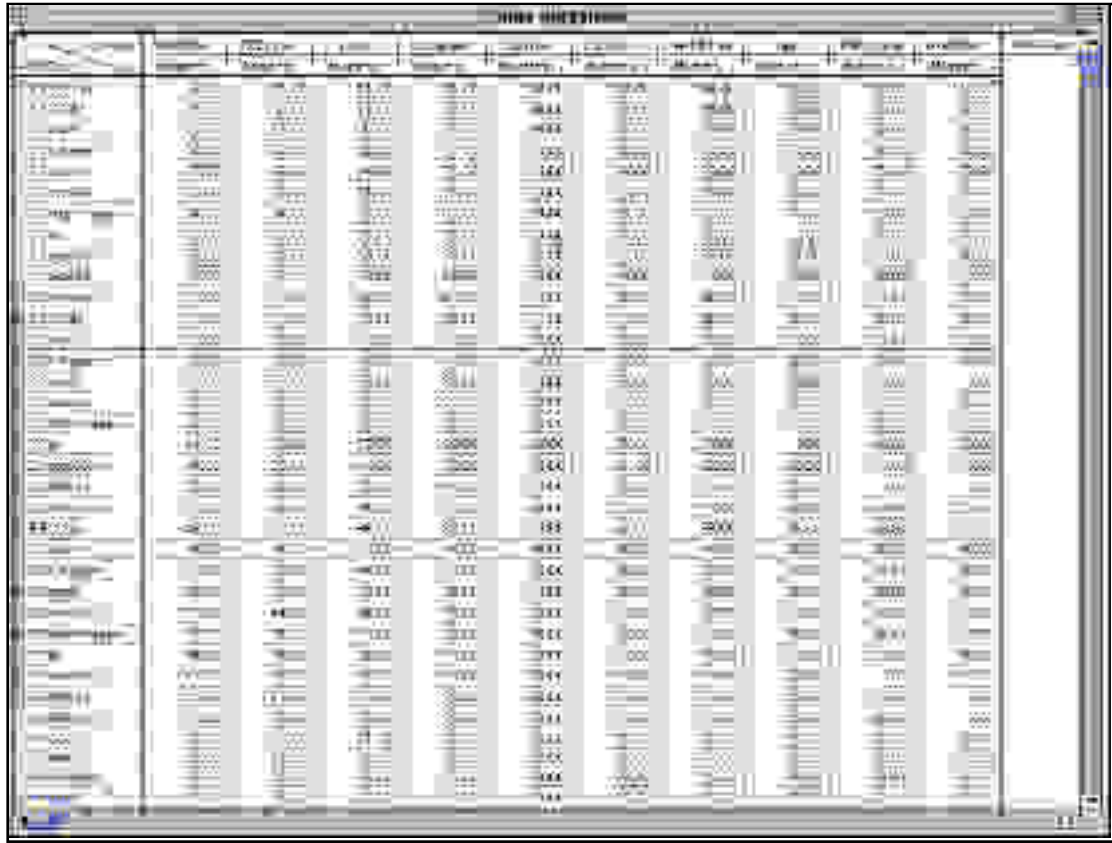


Figura 1.15. Representación de matrices de datos para diferentes sujetos

En la figura 1.17 se muestra la información anterior organizada de una manera más eficiente. Para leerla veremos que en la tabla de la izquierda se encuentra la edad de cada cabeza de familia y un identificador. Ese identificador se utiliza en la segunda tabla para conectar la información de cada padre con la de sus hijos. Así, el cabeza de familia 1 tiene un hijo con 5 años tal y como se indicaba en la tabla anterior. Esta estructura no tiene celdas vacías, y dado un sistema de gestión de datos apropiado, puede resultar más cómodo de utilizar que la tabla anterior. No obstante, como vemos, la representación por medio de una tabla única se ha perdido.

CF	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
50	5	.	.	.	.	.	.	.	.	.
30	3	1	.	.	.	.	.	.	.	.
28	.	.	.	.	.	.	.	.	.	.
45	10	5	.	.	.	.	.	.	.	.
38	12	5	.	.	.	.	.	.	.	.
39	11	5	2	.	.	.	.	.	.	.

Tabla 1.16. Datos organizados en forma de tabla con estructura relacional

Aunque volveremos más adelante a este tipo de representaciones es necesario anticipar algunos de sus inconvenientes. En primer lugar, aunque algunos paquetes estadísticos suelen ofrecer mecanismos para manejar este tipo de datos, es habitual que se produzca un aumento de la complejidad de uso que puede hacer más difícil las tareas de gestión de datos (lo cual lleva a que el ICPSR por ejemplo no recomiende su utilización salvo en casos excepcionales). En segundo lugar, puesto que la mayoría de las técnicas estadísticas están diseñadas para trabajar con datos de forma rectangular, en la práctica los usuarios están obligados a recuperar la tabla rectangular de la figura 1.17 mediante manipulaciones de las tablas relacionales antes de poder llevar a cabo los análisis.

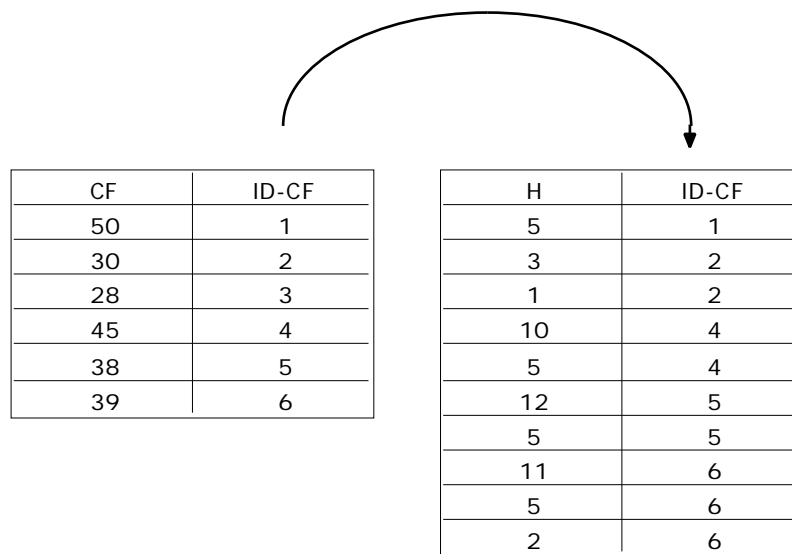


Figura 1.17: Representación de datos con estructuras relacionales

### 1.6.2. Estructura de datos en lenguajes de programación estadística.

Esta descripción está basada sobre todo en las estructuras disponibles en Lisp-Stat (Tierney, 1989; Tierney, 1990) y en S-Plus (Data Analysis Products Division, 1997), los cuales constituyen los sistemas con un repertorio de estructuras diseñadas específicamente para la programación estadística. Un concepto previo es el de tipo de datos. S-PLUS, por ejemplo, maneja los siguientes:

- Lógico. Con valores T (true-verdadero) y F (falso). Este tipo de datos a veces es llamado binario y se representa por medio de 1 y 0.
- Numérico: Números reales almacenados en doble precisión. Pueden ser enteros, con decimales o estar especificados en notación científica.

- Complejos: Números de la forma  $a + bi$ .
- Caracteres: Texto encerrado por comillas tal y como "idea" o 'verano'.

Las estructuras de datos son agrupaciones de datos del mismo o de diferentes tipos. S-Plus hace una distinción entre estructuras de datos que admiten sólo datos del mismo tipo y otras que admiten de diferente tipo. Lisp-Stat no hace esa diferenciación y todas las estructuras admiten datos de diferente tipo.

- Vectores: Los vectores son conjuntos de datos numéricos con sólo una dimensión. Esto significa que la longitud del conjunto de datos puede expresarse mediante un único valor. En S-Plus, más allá de lo que es visible para el usuario, todo son vectores (Data Analysis Products Division, 1997).

- Matrices: Las matrices son conjuntos de datos de dos dimensiones. Su representación visual es por tanto la de una tabla. En realidad, este nombre hace referencia únicamente a un caso particular de la estructura que nombraremos a continuación.

- Matrices multidimensionales (*arrays*). Las matrices multidimensionales son conjuntos de datos con un número de dimensiones mayor que dos. Mediante ellas es posible representar datos que tengan más de dos vías de modo que es posible utilizar organizaciones como cubos o estructuras más complicadas para encajar nuestros datos (Klinke, 1997).

Las estructuras anteriores tienen un gran paralelismo con respecto a operaciones matemáticas. Este paralelismo se pierde en las siguientes estructuras.

- Listas: Una lista es una colección arbitraria de estructuras de datos elementales. Por ejemplo, una lista puede estar compuesta de tres elementos que son números. O pueden ser una mezcla de números y letras. O, más interesante todavía, puede componerse de un vector, una matriz y un array. Naturalmente, una lista puede también estar compuesta a su vez de otras listas.

N	Med1	Med2	Med3
Med1	Var1	Cov21	Cov31
Med2	Cov12	Var2	Cov32
Med3	Cov13	Cov23	Var3

Figura 1.18. Resumen de una matriz de datos multivariada

Entre las ventajas de las listas frente a las estructuras de datos anteriores tenemos que nos permiten poner juntas cosas que poseen longitudes diferentes. Por ejemplo, un resumen de una tabla multivariada de datos podría componerse del tamaño de la muestra (un valor), las medias de las variables (un vector), las covarianzas (una matriz triangular simétrica en la que en la diagonal se encuentran las varianzas). Esta información puede ser almacenada en una matriz si no queda más remedio que utilizar esta estructura tal y como se muestra en la figura 1.18, pero una lista compuesta por un elemento, un vector y una matriz resulta más cómoda de manejar desde el punto de vista de la programación.

Las listas son tan importantes para el lenguaje Lisp que incluso su nombre deriva de ellas (Harrison, 1990, Steele, 1984). Lisp es una abreviatura de List-Processing y esta estructura resulta tan interesante que S-Plus y, naturalmente, Lisp-Stat, la incorporan. Una estructura de gran importancia en S-Plus denominada *data frame* (estructura de datos) se define a partir de listas de otros elementos más simples.

- **Objetos de datos.** Un objeto de datos es un tipo de datos diseñado específicamente para la programación orientada al objeto. Una descripción más completa de estos aspectos corresponde al capítulo 8. Aquí veremos simplemente su aplicación en relación con las estructuras de datos. Un objeto posee información en unas localizaciones denominadas *slots* (ranuras). Estos slots pueden ser definidos por el usuario y llenados con estructuras de datos tal y como vectores, matrices y listas.

Los objetos a su vez incorporan métodos que manipulan esos datos. Algunos métodos básicos pueden ser mostrarse a sí mismos, llevar a cabo una representación gráfica apropiada o añadir o eliminar partes de sus datos. Estos métodos tendrían la ventaja de ser obtenidos de otros objetos por lo cual no sería necesario definir para cada caso los mensajes a responder.

La ventaja de los objetos para el manejo de datos es que podemos hacer definiciones específicas de éstos a partir de modificaciones de otras más abstractas que encajen con la forma que tienen los datos que estamos manejando. Por ejemplo, Tierney (1990) define un objeto de datos genérico destinado a almacenar matrices de datos y lo denomina *data-set-proto*. Ese objeto sólo necesitaría los *slots* de datos y de título. Un objeto de datos de series temporales en cambio podría incorporar un origen y un espaciamiento entre las diferentes observaciones (por ejemplo se empezó en 1960 y los datos son cada dos años). Este nuevo objeto de datos podría incorporar *slots* para esos datos y mantener los mismos que el objeto *data-set-proto*. Objetos de datos mucho más precisos podrían ser definidos de manera similar, así como métodos que realizaran operaciones apropiadas para ellos, para datos que no tuvieran un tratamiento adecuado predefinido (por ejemplo, los tipos de

datos secuenciales discutidos en la sección anterior podrían beneficiarse de una descripción de este tipo). El tema de la programación orientada a objeto es muy rico y será tratado con profundidad en la sección dedicada a programación para el procesamiento de datos.

### 1.6.3. Estructura de datos avanzada: estructura relacional.

Los sistemas computerizados diseñados específicamente para las tareas de gestión y almacenamiento de datos son los denominados bases de datos. Estos programas han evolucionado desde modelos más primitivos basados en estructuras jerárquicas o de red (Date, 1993) al modelo relacional, el cual, gracias al trabajo de Codd (Date, 1993, Losilla Vidal, et al., 1997), posee como ventaja fundamental el estar basado en una conceptualización matemática que determina las reglas y principios a seguir para garantizar su buen funcionamiento.

Entre las características fundamentales de una base de datos está la existencia de una separación entre los programas encargados de la gestión física de los datos y aquellos utilizados para responder a las necesidades concretas de la organización que mantiene esa base de datos (Willitts, 1992). En nuestro caso, esta separación se produciría con respecto a las necesidades concretas de un estudio o investigación. Así, el panorama de uso puede concebirse como que existe por un lado una cierta cantidad de recursos destinada a recoger y almacenar datos que son considerados interesantes para la organización o la investigación, y otra cantidad de recursos dedicada a explotar esa información respondiendo a preguntas concretas (estas últimas tareas son denominadas en ocasiones *data mining* -v. *Communications of the ACM. November 1996.*). En otras ocasiones, las preguntas a responder están claramente especificadas en el momento de recoger la información y la utilización de una base de datos es requerida solamente para facilitar estas tareas.

El modelo relacional de bases de datos es el más importante en la actualidad. En este modelo el concepto básico es el de *relación*, el cual corresponde, desde un punto de vista intuitivo, al de tabla. Antes de continuar es conveniente señalar, tal y como hace Date, que el trabajo de Codd fundamentalmente aportó el formalizar conceptos intuitivos del estilo de tabla, y que para ello prefirió cambiar sus nombres a otros más inusuales, por ejemplo *relación*, para de ese modo acentuar las diferencias con respecto al concepto informal. En nuestro caso señalaremos los nombres formales a la vez que los informales dado que en el espacio que dedicaremos resultaría muy difícil detallar las diferencias que existen entre ellos.



En una relación tendríamos una serie de atributos (cuyo número se denomina grado) que corresponderían a las columnas de la tabla, y unas *tuplas* (cuyo número se denomina cardinalidad) y que serían las filas de la tabla.

Un dominio es una colección o conjunto de valores que las tuplas de un atributo pueden tener. Es como un diccionario de los valores admisibles en una columna (o columnas si este diccionario es utilizado por más de una).

Una clave primaria es un identificador único para las filas de una tabla. Es decir, es una columna (o combinación de columnas) con la siguiente propiedad, nunca existen dos filas con el mismo valor en esa columna (o combinación de columnas). La clave primaria garantiza que dos *tuplas* nunca sean iguales.

Las relaciones poseen cuatro propiedades

a) No existen *tuplas* repetidas. Esto se deriva de que el cuerpo de una relación es un conjunto en sentido matemático, y éstos por definición no incluyen elementos repetidos. Esta propiedad nos obliga a tener claves primarias en las relaciones.

b) Las *tuplas* no están ordenadas. En una tabla existe una ordenación de las filas de arriba a abajo. Esto no ocurre en las relaciones.

c) Los atributos no están ordenados. En una tabla existe una ordenación de las filas de izquierda a derecha. Esto no ocurre en las relaciones.

d) Todos los valores de los atributos son atómicos (aquí se asume que hablamos de atributos simples, también hay atributos compuestos que corresponderían a la combinación de esos atributos simples pero éstos se entienden que no tienen una existencia propia sino que son sólo el producto de un cálculo). Esto significa, según Date, que sólo existe un valor del dominio de un atributo en cada tupla. Esta propiedad es la que en términos prácticos probablemente tiene un mayor impacto a la hora de definir una base de datos relacional. Un ejemplo puede ayudar a aclarar su significado. Supongamos que tenemos una base de datos de productos a la venta. Esta base de datos podría incluir el nombre, el tamaño y otras características. Supongamos que uno de los productos viene en dos tamaños (aunque no varía en ninguna de las otras características). La tentación obvia sería introducir ese segundo tamaño junto al primero, sin añadir una nueva tupla. No obstante, esa tupla con dos valores para un atributo rompería el modelo relacional para los datos, haciendo mucho más difícil llevar a cabo las operaciones relacionales tal y como son habitualmente definidas. Este atributo con varios valores se denomina grupo de repetición.

Las propiedades anteriores, aunque necesarias para poder utilizar el álgebra relacional (las cuales derivan de las operaciones matemáticas sobre conjuntos y son interesantes por su claridad y simplicidad), presentan el inconveniente de introducir dos efectos indeseables. Estos son la duplicidad y la redundancia de la información.

- **Duplicidad:** La información está duplicada en una tabla cuando dos tuplas o más de una tabla contienen valores repetidos en un atributo. Por ejemplo, si dos sujetos han nacido en la misma ciudad, el nombre de la ciudad estará duplicado.

- **Redundancia:** La información es redundante cuando se ha introducido información en una tupla que está duplicada con respecto a otra (lo cual coincide con la idea de duplicidad) que es deducible a partir del resto de la información en la tupla. Por ejemplo, en los datos de envío a una empresa siempre figura el contenido del envío y la dirección de la empresa. Puesto que la dirección es siempre la misma, ésta podría ser deducida a partir del nombre de la empresa, por lo que duplicarla implica introducir redundancia.

La repetición de la información es un inconveniente por dos razones:

- Desde el punto de vista del almacenamiento de la información se está produciendo un desperdicio del espacio ya que hay un aumento en el espacio físico necesario para registrar los datos. Esta consideración, no obstante, no es la más crítica para llevarnos a eliminar estas repeticiones siendo la siguiente consideración más importante.

- Las repeticiones plantean serios problemas desde el punto de vista de la consistencia de los datos. Supongamos que la empresa del ejemplo anterior cambia de dirección. Para poner al día la base de datos necesitaríamos buscar todos los envíos que esa empresa ha realizado y cambiar, en todos ellos, la dirección allí incluida. Esto puede realizarse en una base de datos con una estructura sencilla, pero cuando la información está repetida en diversos archivos o lugares lo más probable es que algunas de las fuentes se queden sin ser actualizadas, por lo que la información pasará a ser inconsistente con los consiguientes inconvenientes.

El proceso mediante el que se elimina la información duplicada y/o redundante es denominado normalización y será descrito a continuación.

*Normalización* es una técnica de modelización de datos descrita por Codd en los años 70 para determinar las propiedades de una relación estructurada correctamente. Los conceptos básicos para llevar a cabo una normalización son los siguientes (Willitts, 1992).

- a) La existencia de una clave primaria o identificador.

b) Los grupos de repetición

c) Las dependencias funcionales entre atributos en una tabla.

Puesto que el concepto de clave primaria y el de grupos de repetición ha sido expuesto anteriormente pasaremos directamente al de dependencia funcional.

Se dice que un atributo B es funcionalmente dependiente de un atributo A cuando, dados dos valores en A,  $a_1$  y  $a_2$ , y otros dos en B,  $b_1$  y  $b_2$ , es verdad que:

$$\text{Si } a_1=a_2 \text{ , entonces } b_1=b_2 .$$

En otras palabras, A es determinante de B si cada valor de A tiene un valor asociado en B.

Algunos ejemplos de determinación funcional simple pueden ser la provincia y la comunidad autónoma, un producto y la empresa que lo fabrica (en caso de que sólo haya una fábrica). Un caso de determinación funcional en que B es determinado por una combinación de atributos son las asignaturas cursadas y el curso que actualmente se está siguiendo (sólo una asignatura no determina el curso que se está siguiendo).

Un concepto importante es el del conjunto mínimo de atributos que es capaz de determinar funcionalmente el resto de los atributos en la relación. Este conjunto mínimo es un identificador o clave primaria ya que éste atributo determina los valores en el resto de los atributos. Por ejemplo, el sujeto con un número de DNI dado tiene consecuentemente una serie de atributos como son su ciudad y fecha de nacimiento, etc.

*El proceso de normalización* convierte una relación con información redundante en varias relaciones que poseen una estructura más simple. Esto se hace siguiendo el principio de buscar tener un sólo valor en cada sitio, de modo que las dependencias funcionales entre los atributos no primarios sean las menos posibles. Ello hace que operaciones básicas en las bases de datos tal y como seleccionar, actualizar o borrar tuplas sean posibles de una manera simple, sin necesitar examinar las posibles dependencias o repeticiones de la información.

Este proceso se basa en tres fases, las cuales producen lo que se denominan formas normales (en realidad hay más de tres formas normales, pero éstas son consideradas suficientes en la mayoría de los casos prácticos). Cada una de estas formas normales es un estado de la base de datos que cumple una serie de condiciones cada vez más estrictas, las cuales son progresivamente más interesantes para la gestión de la base de datos.

La *primera forma normal* (1FN) se da para una relación cuando todos los dominios subyacentes contienen sólo valores atómicos.

Puesto que no puede haber filas repetidas en esta forma normal debe existir una clave primaria (ya sea compuesta de un atributo o una combinación de ellos). Esta relación a veces se denomina "universal" porque incluye toda la información en la base de datos. No obstante, esta forma puede tener una gran cantidad de información duplicada. En el ejemplo de la figura 1.19 nos encontramos con una clave primaria (el ID - Estudiante) que garantiza que ningún registro estará duplicado porque le asignaremos a cada estudiante un valor diferente (aunque su nombre pueda coincidir con el de otro estudiante). Esta tabla presenta varias dependencias funcionales con respecto a la información del colegio. En primer lugar vemos que el colegio A tiene actividades extraescolares y el B no. En segundo lugar vemos que el colegio A tiene en el curso 1 una optativa que el colegio B tiene en el curso 2. Esas dependencias son indeseables y dan pie al planteamiento de la segunda forma normal.

### 1FN

ID-Estudiante	Nombre	Colegio	Curso	Extraesco	Optativa
1	Juan	A	1	Si	Si
2	María	A	2	Si	No
3	Antonio	A	1	Si	Si
4	Pedro	B	1	No	No
5	Eva	B	2	No	Si
6	Luis	C	3	Si	Si

Figura 1.19. Relación en primera forma normal

La *segunda forma normal* (2FN) se da para una relación si ésta está en 1FN y todos los atributos no clave dependen por completo de la clave primaria.

Para pasar de una forma a otra llevaremos a cabo una descomposición de la relación en otras relaciones que ilustraremos de modo informal con el ejemplo anterior. El lector interesado puede consultar Losilla Vidal, et al. (1997) para examinar ejemplos más detallados. La descomposición de la relación anterior nos produce dos tablas que estarían en la segunda forma normal tal y como se muestran en la figura 1.20. En la primera tabla todos los atributos dependen del ID-Estudiante, mientras que en la segunda la actividad extraescolar depende del colegio (los colegios A y C tienen actividades extraescolares, el B no). Si queremos saber si un estudiante tiene actividades extraescolares podemos ver el colegio en el que está inscrito y si ese colegio ofrece las actividades extraescolares. Incluso en un ejemplo tan pequeño podemos ver las ventajas de este cambio en la

estructura de los datos. Supongamos que el colegio B pasa a ofrecer actividades extraescolares. Esto supondría un cambio de un sólo valor en la segunda tabla, y la primera no necesitaría ningún cambio, estando la información actualizada automáticamente.

## 2FN

ID-Estudiente	Nombre	Colegio	Curso	Optativa
1	Juan	A	1	Si
2	María	A	2	No
3	Antonio	A	1	Si
4	Pedro	B	1	No
5	Eva	B	2	Si
6	Luis	C	1	Si

Colegio	Extraesco
A	Si
B	No
C	Si

Figura 1.20. Relación en segunda forma normal

No obstante, la primera relación todavía incluye una cierta dependencia, producida por el atributo optativa. Este atributo depende del colegio y el curso en el que esté inscrito el alumno.

La *tercera forma normal* (3F) se da en una relación cuando está en 2F y todos los atributos no clave no dependen de manera transitiva (es decir a través de otro atributo o por combinación de otros atributos) de otros atributos que no sean la clave primaria.

En nuestro caso el atributo optativa depende de manera transitiva del colegio y del curso. La descomposición mostrada en la figura 1.21 corrige este problema.

La tercera tabla indica en qué curso ofrece cada colegio la optativa. En este caso, la clave primaria de esta relación es la combinación de los dos primeros atributos.

La forma 3FN no es absolutamente satisfactoria (Date, 1993) y Codd definió una nueva versión que hoy en día es conocida por forma normal Boyce/Codd. En la práctica, 3FN cubre la mayoría de las situaciones y, dados los objetivos de este texto, consideramos suficiente lo aquí expuesto.

### 3FN

ID-Estudiante	Nombre	Colegio	Curso
1	Juan	A	1
2	Maria	A	2
3	Antonio	A	1
4	Pedro	B	1
5	Eva	B	2
6	Luis	C	1

Colegio	Curso	Optativa
A	1	Si
A	2	No
B	1	No
B	2	Si
C	1	Si

Colegio	Extraesco
A	Si
B	No
C	Si

*Figura 1.21. Relación en tercera forma normal.*







# ***Preparación y Codificación de Datos***

## **2.1. Introducción**

En esta sección describiremos una serie de cuestiones acerca de la forma en que unos datos provenientes de un estudio deben ser organizados previamente a su análisis. Idealmente, este tipo de cuestiones deberían ser resueltas por el equipo de investigación antes de comenzar la fase real de recogida de datos, aunque muy a menudo se relegan hasta que los datos han sido ya recogidos y se está a punto para su conversión en ficheros electrónicos. Muchos investigadores no anticipan las dificultades inherentes a la gestión de muchas variables y en consecuencia pueden producir errores en sus estudios como resultado de equivocaciones en, por ejemplo, nombrar variables o confundir archivos (Freedland y Carney, 1992). La Guía para la Preparación de datos en Ciencias Sociales del Consorcio Interuniversitario para Investigaciones Sociales y Políticas (ICPSR, 1997), el cual incluye instituciones educativas del mundo entero, señala que la existencia de un plan de Gestión de Datos es crítico para el éxito de un proyecto de investigación ya que el costo de éste puede reducirse significativamente si una serie de cuestiones son planeadas previamente. En concreto, el ICPSR señala los siguientes puntos como más relevantes:

1. Documentación de la introducción y gestión de datos: Del mismo modo que se realizan planes de análisis de datos o de construcción de cuestionarios resultaría interesante

planear la gestión y el archivado posterior de datos. Las siguientes cuestiones deberían ser consideradas:

- a) Estructura de archivos. Esto incluye cuestiones acerca de archivo/s de dato/s, la unidad de análisis, la organización de las fases sucesivas en estudios de panel, etc.
  - b) Convenciones para los nombres. El problema fundamental aquí es el nombre de las variables y la lógica que van a seguir.
  - c) Integridad de los datos. Cómo se convertirán los datos en formato electrónico, qué comprobaciones se realizarán para encontrar valores incorrectos, respuestas inconsistentes, registros incompletos, etc. Este tipo de cuestiones son tratadas en este texto de manera bastante exhaustiva con apartados propios.
  - d) Preparación de un libro de códigos. El libro de códigos explica las abreviaciones y los significados de la información introducida en el archivo de datos.
  - e) Construcción de nuevas variables. Muy a menudo, los datos originales son utilizados para construir nuevas variables que no tienen un referente concreto en el cuestionario original. Este tipo de transformaciones deberían ser planeadas por anticipado y documentadas una vez realizadas.
  - f) Integración. Hasta qué punto pueden combinarse todas las tareas anteriores en un único proceso, de tal modo que sólo sea necesario utilizar un programa de ordenador o un conjunto de programas integrados de tal manera que todo el proceso quede registrado y, en caso de considerarse necesario, se pueda repetir para comprobar su corrección y/o modificar en caso de considerarse conveniente.
2. Estudios piloto de introducción de datos y documentación. Del mismo modo que se realizan estudios piloto para comprobar la corrección y buenos resultados del cuestionario en un estudio de encuesta puede resultar interesante llevar a cabo igualmente una prueba piloto de introducción de datos y de la documentación elegida. Llevar a cabo algunos de los análisis planeados puede resultar también interesante. Los datos utilizados deberían ser lo más realistas posibles ya que en caso contrario se corre el riesgo de no haber previsto situaciones inesperadas que, más tarde, cuando el estudio ha comenzado realmente, tiene consecuencias indeseadas. Estos estudios piloto también permiten hacer estimaciones del coste total del trabajo.
3. Diseño del cuestionario para una buena introducción de datos. En estudios de encuesta el cuestionario es la herramienta principal y una gran parte del esfuerzo del trabajo

consistiría en convertirlo a un formato analizable. Debido a la importancia de este tema será tratado en una sección propia a continuación.

## **2.2. Diseño del cuestionario**

Aunque el diseño del cuestionario puede considerarse que entra dentro de la parte sustantiva de una investigación y por tanto no corresponde ser tratada en un texto de índole metodológica no cabe duda que tener una concepción general de su estructura ayudará mucho al encargado de gestionar los datos.

Bourque y Clark (1992) exponen las características generales que suelen tener los cuestionarios utilizados para recoger la información.

### *a) Número de identificación.*

Todos los cuestionarios deben tener un número de identificación único. Los nombres propios no deberían ser usados porque pueden plantear problemas de confidencialidad. Lo más sencillo es poner un número consecutivo a cada cuestionario a medida que va siendo recogido. Sin embargo, resulta también interesante guardar información acerca del subconjunto de la muestra del que provienen esos datos. Por ejemplo, dos números pueden ser reservados para indicar los colegios a los que se ha entrevistado, otros dos al aula y, finalmente, los últimos dos al alumno. Así, 010205 indicaría el colegio número 1, aula número 2 y alumno número 5.

### *b) Datos a recoger.*

Un suceso bastante desafortunado en un estudio que se encuentra en fase avanzada es descubrir que hay cierta información que no ha sido recogida. Bourque y Clark (1992) recomiendan tener el siguiente esquema como una guía general a considerar a la hora de determinar la información que se desea recoger.

1. Datos demográficos. Esto suele incluir género, edad, nivel educativo, ingresos, profesión, estado civil, etc. Algunas de estas preguntas plantean problemas muy particulares.

- Ingresos. Esta pregunta puede generar muchas no-respuestas puesto que los sujetos a menudo no gustan de dar información acerca de esta cuestión. Las no-respuestas deberían recibir un tratamiento adecuado para evitar resultados equivocados. Una alternativa es preguntar si los ingresos se encuentran dentro de ciertas categorías.

Otra es hacer preguntas acerca de los bienes o servicios de los que disfruta para, a partir de esa información, hacer estimaciones del nivel económico del entrevistado.

- **Profesión.** El número de profesiones existente es ciertamente grande y las implicaciones que tiene el tener una u otra resultan difíciles de delimitar. A menudo, el objetivo indirecto de preguntar por las profesiones es hacer una clasificación del nivel socio-económico del entrevistado, lo cual incluiría por ejemplo información acerca del nivel educativo, etc. Es decir, aunque la información acerca de las profesiones es fundamentalmente categorial los investigadores intentan darle algunas cualidades ordinales. En términos de la clasificación de Tukey (1977) se desea considerar esta variable como de tipo grado o de nivel. No obstante, para que esta información resulte útil es necesario plantear los objetivos deseados con claridad y elaborar una categorización de las profesiones adecuada para satisfacerlos. Esto puede realizarse antes de recoger los datos o, de un modo más costoso, posteriormente, planteando la pregunta como de tipo abierto y categorizándola tal y como se explicará más adelante. Otra solución es utilizar una clasificación de profesiones tal y como la utilizada en estudios de censo (Bourque y Clark, 1992).

- **Nivel de estudios.** De nuevo la categorización planteada suele establecerse en términos de imponer un cierto orden. No obstante, muchos tipos de estudios suponen problemas para esa ordenación (¿un estudiante de Formación Profesional es más o menos que un estudiante de Bachillerato?). De nuevo, una atención cuidadosa a los objetivos planteados puede resultar la mejor forma de resolver el problema.

2. Información propia del estudio. Bourque y Clarke (1992) señalan las siguientes áreas: información acerca del contexto del entrevistado, conductas, experiencia o status y pensamientos o sentimientos. A menudo dos o más de estas áreas serán importantes en un estudio. Por ejemplo, conocer la ciudad de residencia y el área concreta dentro de ésta en la que vive un entrevistado puede ayudar a entender sus opiniones acerca de determinados temas. Una descripción de los tipos de preguntas que pueden realizarse en un estudio van más allá de los propósitos de este texto, pero el lector interesado puede consultar Saris (1998) para una clasificación más detallada de preguntas y Krosnick y Fabrigar (1996) para consideraciones acerca del efecto de la formulación de las preguntas en cuestionarios tal y como por ejemplo la utilización de preguntas abiertas o cerradas (capítulo 2), usar jerarquización (ranking) o valoración (rating-capítulo 3), el número de puntos a utilizar en la escala (capítulo 4), el orden de las alternativas de respuesta (capítulo 7) y el problema de la aquiescencia o tendencia a decir "sí" siempre en cuestionarios (capítulo 8).

### **2.2.1. La forma de la pregunta.**

Un elemento de gran importancia para el procesamiento de datos estriba en la utilización de preguntas abiertas o cerradas en el cuestionario. En las preguntas cerradas existe una lista de respuestas que el entrevistado debe contestar. En las preguntas abiertas no existe esa lista y el entrevistado tiene la oportunidad de producir una contestación que, potencialmente, podría ser larga y compleja (Bourque y Clarke, 1992).

En general, las preguntas cerradas ofrecen una información más fácilmente resumible por medio de análisis estadísticos por lo que en estudios que impliquen gran número de entrevistados es usualmente la forma más cómoda de recoger la información en muchos de los casos. No obstante, a menudo ocurre que ciertas cuestiones sólo pueden ser tratadas utilizando preguntas en las que el entrevistado no se encuentra limitado a una serie de respuestas prefijadas. Este tipo de preguntas idealmente no debería constituir una parte muy importante del cuestionario pues supone hacer muy complejo el procesamiento de datos. Existe software sin embargo que contempla como objetivo principal la gestión de información de este tipo y que será comentado.

Las preguntas abiertas poseen la ventaja de que no es necesario anticipar todos las posibles contestaciones de los entrevistados por lo que resultan más fácil de construir. Sin embargo, presentan el inconveniente de alargar el tiempo necesario para realizar la entrevista. Una forma de mejorar las preguntas cerradas sería realizar estudios previos con preguntas abiertas para determinar todas las posibles contestaciones que los sujetos podrían dar. Eso haría, contrariamente a lo presupuesto normalmente, que utilizar preguntas cerradas fuera más costoso que utilizarlas abiertas. Utilizar preguntas cerradas con una opción de "otras" para alternativas no indicadas no resulta muy útil ya que el efecto de primacía hace que los sujetos se inclinen por aquellas mencionadas explícitamente (Krosnick y Fabrigar, 1995).

### **2.2.2. Preguntas cerradas.**

La elaboración de preguntas cerradas debería seguir las siguientes reglas según Bourque y Clarke (1992)

1. Exhaustividad. Las categorías de respuesta deberían representar el espectro de todas las posibles respuestas. En caso de que esto no sea posible, o sea especialmente difícil y/o problemático debería incluirse una categoría residual del tipo "Otros" o similar. Estas otras contestaciones podrían ser escritas individualmente. El uso de esta categoría residual no impide otro problema: los sujetos pueden optar por categorías listadas antes

que producir las propias estando los resultados muy limitados en caso que no se elijan bien las alternativas (efecto de primacía). Dos efectos que pueden producir unas respuestas que cubran mal las posibles respuestas son:

- El apilamiento o respuestas concentradas en una única categoría. Por ejemplo se utilizan como categorías de respuestas para la pregunta salario en pesetas percibido al mes por el individuo las siguientes alternativas: 1. Menos de 30.000; 2. Entre 30.000 y 300.000; 3. Más de 300.000. En este caso, se producirá bastante obviamente un apilamiento en la alternativa 2.
- Datos no recogidos. No incluir una alternativa puede llevar a los entrevistados a contestar cualquier otra que se les ajusta peor pero está disponible.

2. Alternativas Excluyentes. Las alternativas deberían en principio ser mutuamente excluyentes entre sí. Muchos análisis estadísticos parten de este supuesto y, en cualquier caso, la interpretación de información solapada entre sí puede ser muy confusa.

3. Alternativas fáciles de comprender. Las dos recomendaciones anteriores pueden estar en conflicto con la elaboración de alternativas fáciles de comprender por el respondente o incluso por el entrevistador. Esto hará que el momento de realizar la encuesta se alargue en exceso o que el entrevistado opte por una alternativa cualquiera o bien cese su participación.

4. Categorías residuales. Como se ha señalado anteriormente, una forma de ser exhaustivos es añadir a cada pregunta una categoría de "otras". Esto debería hacerse incluso en el caso que se crea que todas las opciones han sido contempladas.

5. Anticipar necesidades de análisis. En muchas ocasiones se plantean preguntas utilizando alternativas cuando una que puede ser contestada de modo abierto sin ninguna ambigüedad (por ejemplo, la edad) produce mejores resultados. Por ejemplo, la mayor parte de las veces tiene más sentido preguntar "¿Cuántos años tiene usted?" que "Dígame si tiene usted entre 20 a 30 años, entre 30 a 40, o entre 40 a 50". En cualquier caso, si el investigador lo que desea es el segundo tipo de información siempre puede utilizar un paquete estadístico para recodificar la información.

6. Evitar información faltante. Cierta tipo de preguntas produce altas tasas de rechazo y de no contestación. Es conveniente por tanto plantearlas de tal modo que se evite este efecto. Una posibilidad es por ejemplo hacer lo contrario que en la recomendación hecha en el punto anterior. Supongamos que estamos interesados en el salario de los individuos, pero anticipamos que seguramente se sentirán incómodos en

caso de preguntárselo directamente. Hacerlo utilizando categorías puede, en cambio, resultarles más aceptable de responder que proporcionar una cifra exacta. También se puede obtener información a través de indicadores indirectos. Saris y Gallhofer (1998) indican el siguiente ejemplo. La pregunta "¿Está usted en el censo electoral de esta zona?" puede ser utilizada en lugar de "¿Tiene usted derecho a votar en este país?".

Finalmente, es necesario recalcar que para evitar sorpresas resulta muy recomendable realizar pruebas piloto de las preguntas de los cuestionarios con una muestra reducida para evaluar su funcionamiento.

### **2.2.3. Preguntas abiertas**

En una pregunta abierta, el respondente no tiene una lista cerrada de alternativas a las que contestar. Así, por ejemplo, la siguiente pregunta sería abierta.

- ¿Qué marcas de desodorante conoce usted?

Mientras que la siguiente sería cerrada.

- De esta lista de desodorantes (Sanex, Tulipán Negro, Fa, Tacto, Nívea, Vasenol, Otros)...¿cuáles conoce usted?

Ahora bien, existe una cierta progresión en cuanto a lo "abierto" que puede ser una pregunta, dando lugar a situaciones que podríamos considerar progresivamente más complejas. Estas serían:

a) Cuando sólo es posible contestar con una cifra o valor, por ejemplo, al preguntar la edad, o cuando la lista de alternativas es conocida implícitamente por los respondentes, por ejemplo, cuando se pregunta por el estado civil. Este caso no es considerado verdaderamente una pregunta abierta. De todos modos, tal y como señalan Bourque y Clarke (1992) resulta habitual que los entrevistados den respuestas que no se ajusten al formato esperado por lo que es posible que sea necesario un reajuste posterior de las contestaciones. Por ejemplo, a la pregunta, "¿cuántas personas estaban presentes cuando se produjo el accidente?", la contestación esperada sería un número, tal y como 3 ó 5, pero mucha gente contestará "varios" o "no lo sé". Este tipo de contestaciones alternativas deberían ser previstas y se tendría que ofrecer una alternativa.

b) La contestación se limita a una sola palabra o una lista de palabras. Por ejemplo, preguntar por la marca de desodorante que usa el entrevistado está limitado a una sola palabra. Hacerlo por marcas de desodorante conocidas o utilizadas en alguna ocasión aumenta las posibles contestaciones a una lista más amplia.

c) La contestación es del tamaño de un párrafo aproximadamente y la información en ella contenida puede ser resumida por medio de la técnica conocida como análisis de contenidos. Por ejemplo, preguntar la opinión acerca de los servicios en un área urbana.

d) La información consiste en textos de longitud bastante grande, posiblemente interrelacionados, y sobre los que se desean hacer estudios y comparaciones de tipo cualitativo antes que cuantitativo. Por ejemplo, un estudio sobre artículos publicados en periódicos acerca de un determinado acontecimiento.

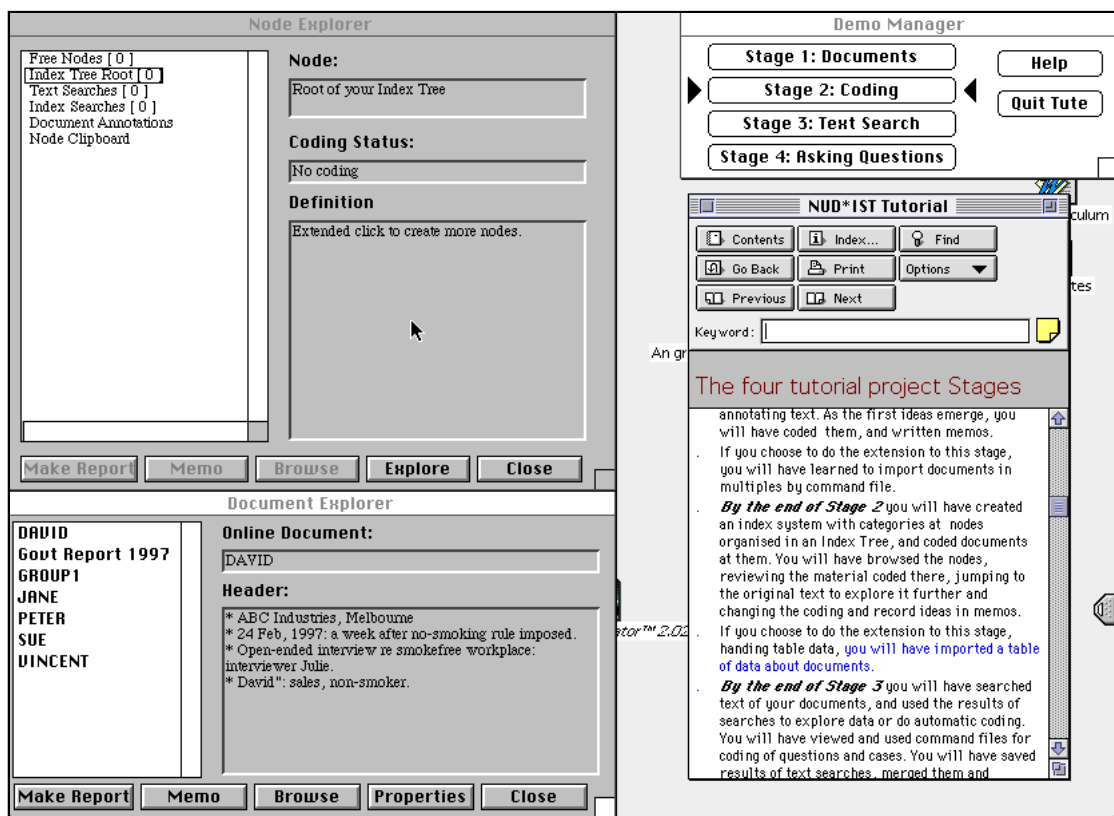


Figura 2.1. Pantalla de NUDIST (non-numerical unstructured data indexing, search and theorizing)

Naturalmente, cuanto más simple sea la situación menos problemática será. En este texto nos centraremos en las tres primeras situaciones y omitiremos el último caso, referido a un tipo de investigación centrado en el uso de datos cualitativos, los cuales tienen una forma de palabras antes que de números (Miles and Huberman, 1994), y que han sido los más habituales en disciplinas como Antropología, Historia y Ciencias Políticas, y que actualmente están siendo incorporados también en otras como la Psicología, la Sociología, etc. Existen programas dirigidos específicamente a la



codificación de estos datos así como a la extracción de regularidades y, en muchos casos, a la información numérica (Weitzman and Miles, 1995). Un programa bien conocido y del que se muestra una pantalla en la figura 2.1 es NUDIST (Gil, et al., 1998). Sin embargo, la codificación en este caso no correspondería realmente a tareas de preparación de datos sino que verdaderamente constituyen un elemento esencial del propio análisis, por lo que su consideración detallada en este texto parece inapropiada.

Centrándonos en los dos tipos de situaciones más habituales en cuanto a preguntas abiertas tenemos, las que las contestaciones posibles son una lista de opciones y aquellas que se realiza una descripción de un pequeño párrafo.

1. Cuando las contestaciones posibles corresponden básicamente a listas entonces podemos tomarlas y utilizarlas como las categorías de nuestro estudio y el único problema radica en establecer el diccionario o listado de contestaciones, así como en preveer el número de columnas en la tabla de datos necesarias para este propósito. En el caso más sencillo, tendríamos que sólo es posible una contestación para cada sujeto. Por ejemplo, a la pregunta "¿qué político actual valora usted más?" las contestaciones posibles podrían ser las de la tabla 2.2.

Sujeto	Político	Cod Político
1	Jose M <sup>o</sup> Aznar	1
2	Cristina Almeida	2
3	Felipe González	3
4	Jose M <sup>o</sup> Aznar	1
5	Jordi Pujol	4
6	Abel Matutes	5
7	Jordi Pujol	4
8	F. González	3

Tabla 2.2. Contestaciones posibles a la pregunta "¿qué político actual valora usted más?"

Puesto que leer la lista completa de políticos en un país sería excesivamente largo los entrevistados contestan sin esta ayuda. Para codificar esta información de modo que pudiera ser utilizada en un paquete estadístico normalmente querríamos relacionar cada una de las respuestas con un número para facilitar su manejo (aunque esto no es absolutamente necesario, pues la mayoría de los paquetes estadísticos manejan textos de este tamaño, resulta conveniente hacerlo así). Esto implicaría también corregir errores de escritura e inconsistencias (p.e. Felipe González y F. González). Finalmente, la información introducida en el ordenador aparecería tal y como es mostrada en la tercera columna de la Tabla 2.2.

Un problema más complejo es el que se produce cuando el número de respuestas posibles es mayor de uno. En este caso, la pregunta sería "¿qué dos políticos actuales valora usted más?". Las contestaciones podrían aparecer como en la tabla 2.3, así como las codificaciones correspondientes.

Sujeto	Político 1	Político 2	Cod Político 1	Cod Político 2
1	Jose M <sup>o</sup> Aznar	Cristina Almeida	1	2
2	Cristina Almeida	Jose M <sup>o</sup> Aznar	2	1
3	Felipe González	Jordi Pujol	3	4
4	Jose M <sup>o</sup> Aznar	F. Alvarez Cascos	1	6
5	Jordi Pujol	F. González	4	3
6	Abel Matutes	F. González	5	3
7	Jordi Pujol	Jose M <sup>o</sup> Aznar	4	1
8	F. González	Jordi Pujol	3	4

*Tabla 2.3. Pregunta abierta codificada*

Si el número de contestaciones posible fuera más de uno entonces nos encontraríamos ante una variable de respuesta múltiple. Este tipo de variables será tratado posteriormente de forma más exhaustiva. Las dos últimas columnas de la Tabla 2.3 muestran el aspecto que tendría la tabla utilizada para introducir la información. Este tipo de preguntas necesitan un tratamiento especial en el momento del análisis por lo que, por ejemplo, el programa SPSS incluye comandos específicos para su manipulación.

Todavía un poco más complicado es cuando no se especifica el número de contestaciones admisibles. Por ejemplo, la pregunta "¿qué políticos valora usted más?" daría una tabla similar a la 2.4.

Sujeto	Político 1	Político 2	Político 3	Político 4	Cod Político 1	Cod Político 2	Cod Político 3	Cod Político 4
1	Jose M <sup>o</sup> Aznar	Cristina Almeida	F. Alvarez Cascos		1	2	6	
2	Cristina Almeida				2			
3	Felipe González	Jordi Pujol			3	4		
4	Jose M <sup>o</sup> Aznar	F. Alvarez Cascos	F. González	Jordi Pujol	1	6	3	4
5	Jordi Pujol	F. González			4	3		
6	Abel Matutes				5			
7	Jordi Pujol	Jose M <sup>o</sup> Aznar			4	1		
8	F. González	Jordi Pujol			3	4		

*Tabla 2.4. Pregunta abierta con número de alternativas indeterminado*

En esta tabla ha sido necesario utilizar cuatro columnas para codificar las contestaciones de los sujetos. No obstante, puesto que sólo uno de ellos ha dado cuatro contestaciones la columna correspondiente está vacía salvo en una casilla. Esto supone

una pérdida de espacio y también la necesidad de tomar precauciones en cuanto al análisis. En concreto, las casillas vacías serán tomadas por muchos paquetes estadísticos como información faltante, de tal modo que ciertos análisis estadísticos pueden optar por ignorar la fila entera en la que hay una casilla vacía. También, el número de respuestas no coincidirá con el número de sujetos entrevistados por lo que estadísticos como proporciones o porcentajes pueden resultar difíciles de interpretar.

Una precaución obvia es utilizar una codificación para las casillas vacías que identifique la causa por la que se ha producido. El apartado dedicado a valores faltantes dará más detalles acerca de esta recomendación.

2. Cuando las contestaciones de los sujetos son textos de cierta longitud la codificación resulta más compleja. En ese caso, el investigador deberá leer cada una de las contestaciones y clasificar la información allí contenida en una serie de categorías que ha definido previamente. Dependiendo de la complejidad de las preguntas el conjunto de categorías capaz de cubrir las contestaciones puede resultar relativamente simple o complejo. Este texto está basado en el supuesto que el estudio utiliza básicamente preguntas de tipo cerrado y añade algunas de tipo abierto de poca complejidad. Para estudios en los que la información es fundamentalmente de tipo textual, Weber (1985) presenta una introducción al análisis de contenidos. Aunque la situación contemplada por nosotros sea más sencilla que la contemplada por este autor resulta interesante exponer los pasos descritos por él para crear un esquema de codificación para textos. Estos son:

a) Definir las unidades a registrar. Es posible utilizar las siguientes unidades: (i) cada palabra, (ii) el significado de cada palabra, (iii) cada frase, (iv) temas, (v) párrafos y, (vi) el texto completo. Utilizar unidades más pequeñas y que necesiten poca interpretación por parte del investigador garantizan un mayor acuerdo entre evaluadores. Sin embargo, no atender al sentido de lo dicho puede distorsionar o confundir la información.

b) Definir las categorías a tener en cuenta. Puesto que la información textual o verbal es muy rica es necesario limitarse de algún modo a unas categorías en particular. Bourque y Clarke (1992) señalan las siguientes cosas a tener en cuenta:

- (1) Especificar los objetivos para los que será utilizada la estructura de códigos.
- (2) Mantener un equilibrio entre poco detalle y mucho detalle.

(3) Maximizar la información recogida y la facilidad de codificación. Una pregunta abierta puede dar lugar a varias variables numéricas interesantes.

(4) Registrar la información faltante.

(5) Agrupar categorías relacionadas.

c) Codificar una muestra de las preguntas.

d) Diagnosticar la precisión o fiabilidad. La precisión en este caso hace referencia a la utilización de un ordenador para hacer la codificación y si lleva a cabo la tarea como se pretendía. La fiabilidad correspondería al caso en que la codificación se hace manualmente, y tiene dos posibilidades: Estabilidad (el mismo codificador evalúa el mismo material dos veces y luego se comparan los resultados) y fiabilidad entre codificadores (dos codificadores diferentes evalúan el mismo material y se comparan los resultados).

e) Revisar las reglas de codificación. En caso de que el diagnóstico anterior sea malo. También, aunque el diagnóstico general sea bueno puede haber puntos concretos que lleven a diferencias entre evaluadores y que deberían ser depurados antes de continuar.

f) Volver al paso c. El proceso continúa hasta que la precisión o la fiabilidad conseguidas sean adecuadas.

g) Codificar todo el texto.

h) Diagnosticar la precisión y la fiabilidad para todo el texto. Esto puede ser particularmente importante cuando se utilizan codificadores humanos puesto que muy a menudo la interpretación de las reglas puede ir cambiando a medida que se lleva a cabo la codificación.

### **2.3. Codificación de textos ayudada por ordenador**

Dentro del apartado de codificación y análisis de preguntas abiertas de cuestionario por medio de ordenador comentaremos dos ejemplos: WordStat y TextSmart. El primero de ellos está integrado dentro del paquete estadístico SIMSTAT, un programa de reducido precio, que puede comprarse a través de Internet (existe una demo limitada a un uso de 30 días) y con un precio bastante reducido. TextSmart es un producto de SPSS Inc. (aunque se vende por separado), con precio mayor y con capacidades más completas.

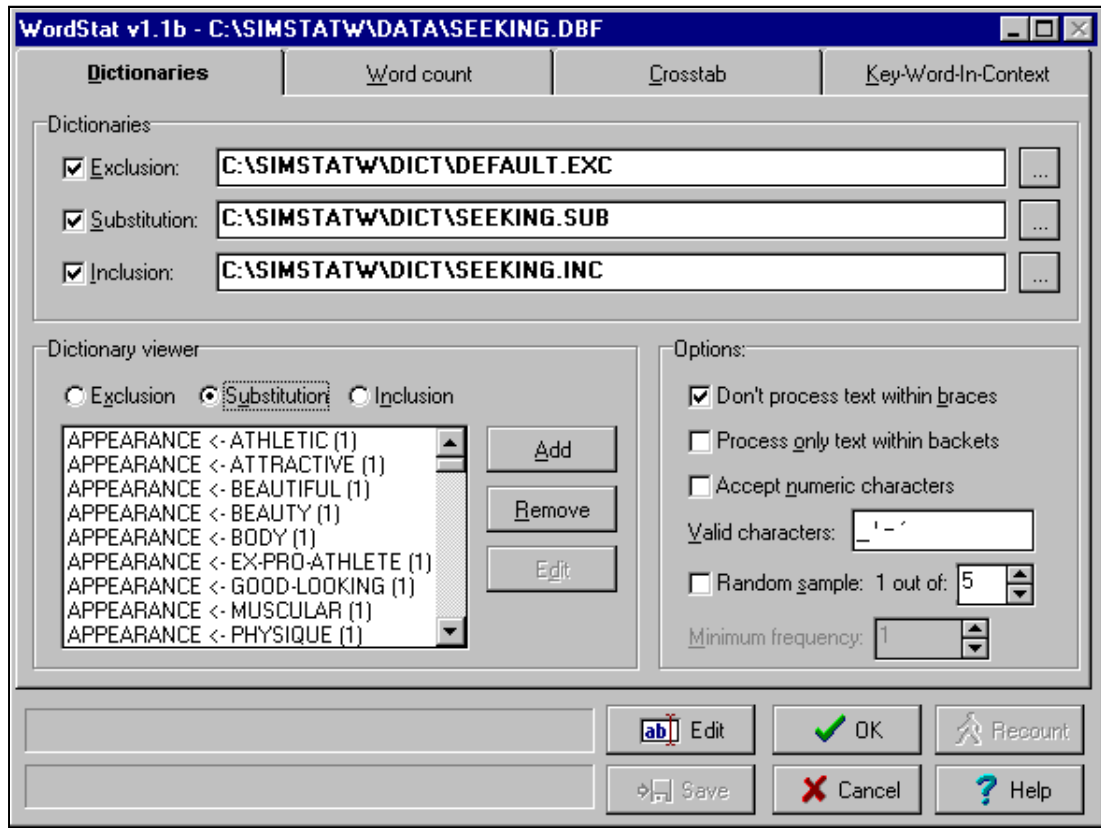


Figura 2.5. Un ejemplo de WordStat.

La secuencia de trabajo habitual puede apreciarse en la figura 2.5 tomada de WordStat. En las pestañas de la parte superior se pueden apreciar cuatro fases en el trabajo con este programa. Estos son:

a) Crear un diccionario de códigos que transformen ciertas palabras en otra palabra que, según el investigador, resume su "tema" o significado. En este caso, el ejemplo utilizado son los anuncios publicados en la sección de contactos de un periódico. Así, tenemos que en el recuadro inferior de la figura 2.5 se observa que los términos ATHLETIC (atlético), ATTRACTIVE (atractivo), BEAUTIFUL (hermoso) etc. serán sustituidos por el término APPEARANCE (apariencia personal). En esa misma sección es posible excluir (EXCLUSION) ciertas palabras por ser consideradas de bajo poder semántico (preposiciones, artículos, etc.) o volverlas a incluir (INCLUSION).

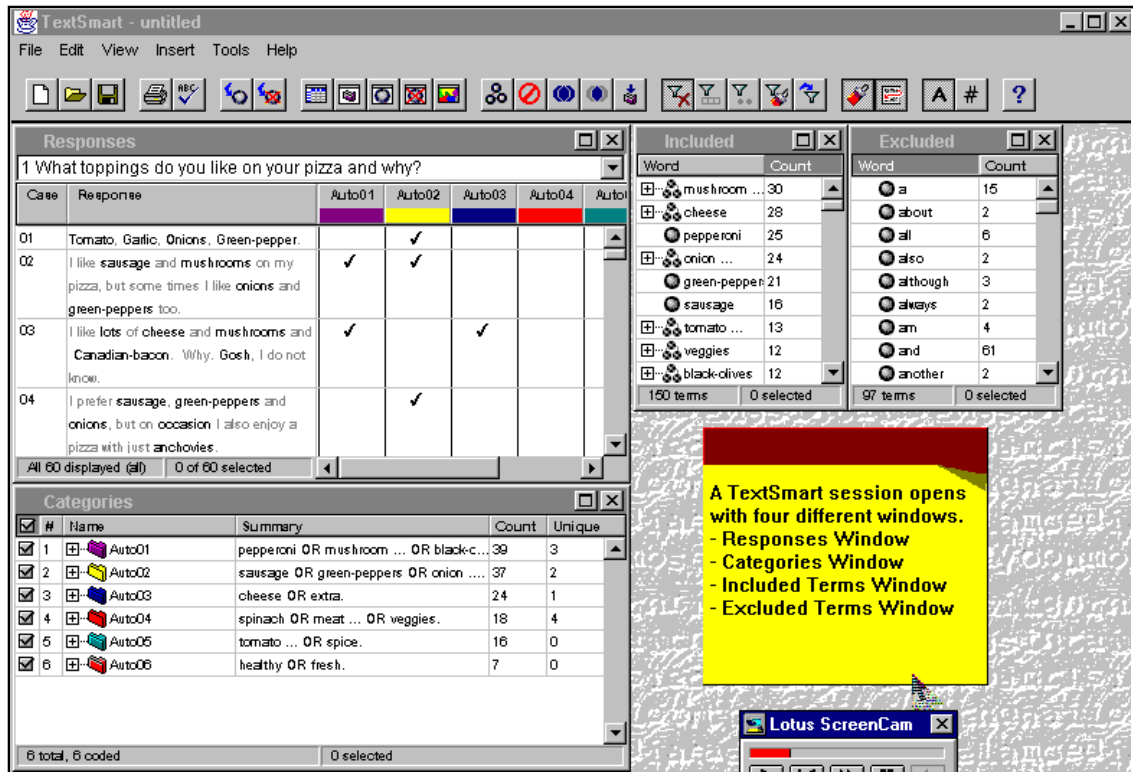


Figura 2.6. Una imagen de TextSmart

En TextSmart es el propio programa el que elabora las sustituciones o categorías. En la parte inferior de la figura 2.5 podemos ver que aparece una ventana con unas categorías anticipadas en las preguntas (en la ventana de la parte superior se ven algunas preguntas de ejemplo). En este caso, el ejemplo versa acerca de los ingredientes que unos clientes gustan en pizzas. El programa ha detectado algunos ingredientes que suelen ir unidos y ha creado las categorías. Por ejemplo, la primera de ellas es Auto01 (el nombre puede cambiarse posteriormente). Incluye salchichón picante (PEPPERONI), champiñones (MUSHROOM) y aceitunas negras (BLACK OLIVES). Alguno de los ingredientes de esta categoría aparece en 39 preguntas (por ejemplo, las preguntas 2 y 3 incluyen champiñones) pero sólo lo hacen de modo único en 3. Esta categorización previa no es probablemente muy perfecta por lo que el investigador podrá perfeccionarla o modificarla a su gusto.

WordStat v1.1b - C:\SIMSTATW\DATA\SEEKING.DBF

Dictionary: Word count Crosstab Key-Word-In-Context

Words to display: Included Sort by: Word frequency List Tree

WORDS	NB WORDS	% SHOWED	% TOTAL	NB RECORDS	% RECORDS
APPEARANCE	83	23.1%	2.6%	43	59.7%
HUMOR	45	12.5%	1.4%	31	43.1%
EDUCATION	40	11.1%	1.2%	23	31.9%
NIGHTLIFE	34	9.4%	1.0%	23	31.9%
ARTS	33	9.2%	1.0%	21	29.2%
COMMUNICATION	27	7.5%	0.8%	18	25.0%
WORK	22	6.1%	0.7%	17	23.6%
OUTDOOR	21	5.8%	0.6%	14	19.4%
SPORTS	17	4.7%	0.5%	12	16.7%
SEXUALITY	16	4.4%	0.5%	13	18.1%
SPIRITUALITY	11	3.1%	0.3%	6	8.3%
FAMILY	6	1.7%	0.2%	5	6.9%
FINANCE	5	1.4%	0.2%	5	6.9%

Edit OK Recount Save Cancel Help

Figura 2.7. Recuentos de las categorías en el ejemplo utilizando WordStat.

b) Recuentos de palabras y análisis. Los dos programas permiten construir tablas de recuentos de las categorías. Estos recuentos pueden ser para cada una de las categorías (v. figura 2.7) o para la combinación de categorías en contestaciones, bien directamente o mediante gráficos obtenidos mediante escalamiento multidimensional (v. figura 2.8).

c) Exportación a otros programas de análisis estadísticos. Ambos programas permiten realizar exportar las categorías y los recuentos producidos a otros paquetes estadísticos o a formatos neutros que permiten su traducción sencilla.

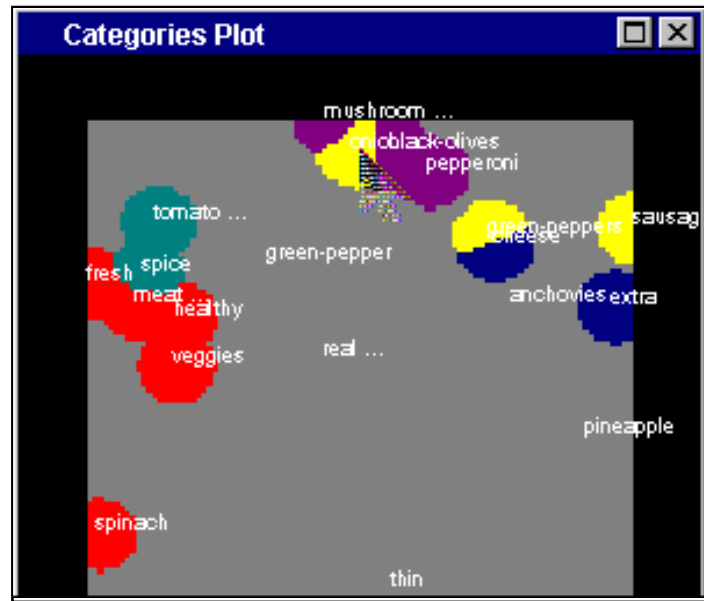


Figura 2.7: Gráfico producido por TextSmart mostrando la configuración de palabras en el diccionario y solapamiento entre ellas.

## 2.4. Diseño del cuestionario para la introducción de datos

Diseñar el cuestionario teniendo en cuenta la introducción y el análisis de los datos puede suponer un ahorro de costos y una disminución de errores. Bourque y Clark (1992) sugieren las siguientes:

1. Utilizar códigos para datos no disponibles o faltantes que no puedan ser confundidos con los otros valores. Una costumbre es utilizar valores negativos formados por nueves (por ejemplo, en una pregunta con códigos del 1 al 6 se indicaría con valor faltante -9 ó -99). Otros paquetes estadísticos permiten utilizar simplemente un punto (.) pero en aquellos casos en que se admitan valores faltantes de varias clases esto puede ser insuficiente. También es importante indicar la causa o razón por la que un valor es faltante. Cada pregunta debería tener alternativas del tipo "No se", "Se niega a contestar", "Inaplicable" o cualquier otra que sea adecuada. Dejar espacio disponible para explicar las causas de la no respuesta puede resultar también interesante. Posteriormente el investigador puede optar por mezclar los códigos, pero esto debería ser una elección tomada durante el análisis y no antes.

2. Preguntas semejantes deberían tener códigos semejantes. Muchas preguntas tienen alternativas semejantes, tal y como "si", "no", "no se". En ese caso es conveniente



utilizar siempre el mismo código. Por ejemplo, "si" podría ser siempre 1 y "no" 0. "No se" podría ser siempre -99.

3. Códigos que minimicen las transformaciones a realizar durante el análisis de datos y que correspondan a significados de la vida cotidiana. Si, por ejemplo, una de las contestaciones es "ninguno", un código natural para esta alternativa es 0. Si se pregunta acerca de dinero los sujetos pueden encontrar más fácil hablar en miles de pesetas antes que en cientos de miles, aunque luego el investigador decida utilizar estos como unidad de medida.

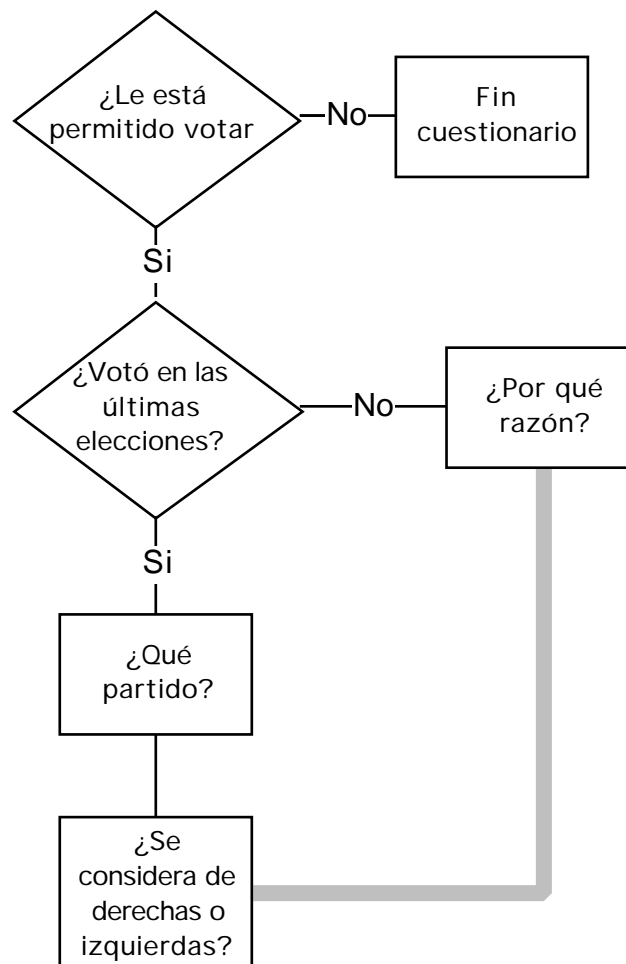


Figura 2.9. Ejemplo de diagrama de flujo para un cuestionario sencillo

4. Patrones de saltos. A menudo no todas las preguntas en un cuestionario son aplicables a un sujeto. Así, dependiendo de contestaciones anteriores ciertas preguntas

serán innecesarias o inaplicables. No obstante, tal y como señalan Bourque y Clark (1992), incluso patrones de saltos bien diseñados pueden llevar a datos incompletos o difíciles de analizar. Por ejemplo, si nos hemos planteado como objetivo recoger datos acerca de  $n$  sujetos, y ciertas preguntas sólo son aplicables al sexo varón, la información obtenida acerca de ellas estará limitada a aproximadamente  $n/2$ . Esto puede resultar un inconveniente importante. Por otro lado, patrones de saltos complejos pueden llevar a que ciertas preguntas sean aplicables en muy pocos casos u otro tipo de consecuencias negativas. Para evitar esas consecuencias resulta muy conveniente realizar diagramas de flujo como el mostrado en la figura 2.9. En él señalamos con un rombo las preguntas que ramifican el cuestionario y con cuadrados aquellas que no. Como es posible ver, la primera pregunta diferencia entre los que votaron y los que no, llevando a continuación a cuestiones diferentes. Este gráfico nos permitiría detectar que la pregunta acerca de si se es de derechas o de izquierdas no se hace a aquellos que no han votado en las últimas elecciones, lo cual, aparentemente, parece una equivocación de diseño del cuestionario. La línea gris corrige esta situación.

## 2.5. Codificación de datos

Las contestaciones verbales de los sujetos deben ser convertidas en códigos numéricos para poder ser manejadas con comodidad en un paquete estadístico (Swift, 1996). Además, la información debe de ajustarse a la estructura en tablas que suele ser habitual en ellos pero sin perder manejabilidad ni inducir a errores. No existen reglas fijas pero es conveniente tener una serie de orientaciones que fomenten la "buena práctica". Dividiremos la presentación en cuestiones acerca de variables, de casos y de valores, aunque en realidad esta distinción es un poco artificial y algunas de las cuestiones podrían haber sido incluidas en apartados diferentes.

- **Aspectos de la codificación a tener en cuenta con las variables.**

Las variables generalmente toman la forma de columnas en la tabla utilizada para representar los conjuntos de datos en paquetes estadísticos. La mayoría de éstos permitirán manejar los nombres que reciben las variables, las etiquetas de las variables y la posición que ocupan. Veamos estos aspectos con más detalle:

a) Nombres de variables. Cuando se está manejando un archivo de datos con un número grande de preguntas el investigador necesita utilizar nombres de variables que le resulten poco ambiguas y transmitan la información correcta en poco espacio. La razón

para la limitación de espacio es que, puesto que estos nombres de variables aparecerán en los resultados que produce el paquete estadístico, es necesario poner un límite o no se verán correctamente. Por esta razón, o por limitaciones internas de los programas, los nombres de variables a menudo no pueden tener más de 8 caracteres. Esta longitud debería bastar si se construye un conjunto de nombres razonablemente organizado. El ICPSR (1997) comenta los siguientes métodos:

1. Usar una serie de números desde 1 hasta  $n$ , en donde  $n$  es el número total de variables. Puesto que muchos programas de ordenador no permiten usar nombres de variables que empiecen por un número lo más común es  $V1 \dots Vn$ . Esta aproximación es simple pero los nombres de las variables no transmiten mucha información y pueden inducir a errores.

Una ventaja de este método es que muchos programas estadísticos permitirán escribir comandos del tipo: <ACCION  $V1$  hasta  $Vn$ > en donde ACCION puede ser cualquier operación estadística como por ejemplo obtener la media. Por ejemplo, esto permitiría pedir la media de las variables desde  $V1$  hasta  $Vn$  sin ser necesario especificar los nombres intermedios.

Una opción es referirse a la pregunta del cuestionario mediante  $C1 \dots Cn$ . Puesto que muchas preguntas producirán varias variables se pueden utilizar letras detrás de los números ( $C1, C2a, C2b, C2c, C3 \dots Cn$ ).

2. Usar abreviaturas. Esto tiene la ventaja de producir información acerca de cada pregunta pero la limitación que las abreviaturas pueden ser difíciles de reconocer para alguien diferente al que las creó (y en ocasiones incluso para el que las creó). Además, ciertas preguntas que tienen información similar pueden ser difíciles de codificar consistentemente entre sí.

<p>Valora de 1 a 5.</p> <p>1. Cuando estoy bien con mis compañeros me siento muy a <b>gusto</b> en la escuela.</p> <p>2. La <b>tensión</b> que me produce la escuela me influye también fuera de ella.</p> <p>3. Me fastidian las <b>reglas</b> escolares y me cuesta respetar alguna de ellas.</p>
---

Figura 2.10. Tres preguntas de un cuestionario referidas a satisfacción escolar

3. Sistemas de prefijos y sufijos. Una forma sistemática de la propuesta 2 es pensar en cada tema como compuesto de un prefijo y un sufijo. Por ejemplo, todo lo referido a

nivel educativo podrían tener el sufijo EDU. El nivel educativo del padre podría ser PA\_EDU y el de la madre MA\_EDU.

El ICPSR parece inclinarse por el método 3. En nuestra opinión, la combinación del método 3 con el 1 puede ser la óptima. Esto es verdaderamente útil cuando se tienen escalas dentro de los cuestionarios en las que varias preguntas hacen referencias a contenidos muy similares. Por ejemplo, en la figura 2.10 aparecen tres preguntas extraídas de un cuestionario de aspectos escolares y referidas a satisfacción con la escuela.

Una codificación posible de los nombres de las variables para estas tres preguntas sería (hay que pensar que en el cuestionario completo se recogía información acerca de otros aspectos además de la satisfacción escolar): GU\_SA, TEN\_SA y REG\_SA. Una opción es 25\_GU\_SA\_1, 25\_TEN\_SA\_2, 25\_REG\_SA\_3. El primer número indicaría el número de variable, luego estaría la información abreviada de la variable y finalmente la pregunta concreta dentro de la escala (no obstante, hemos usado más de ocho caracteres, así que es muy posible que este método no sirva en el paquete estadístico concreto que usemos).

b) Etiquetas de variables. Una forma de solucionar el problema de los nombres de variables con poca longitud, habitualmente encontrados en muchos paquetes estadísticos, es la posibilidad de utilizar etiquetas de variables. Estas etiquetas son textos de cierta longitud que es posible ligar a las variables pero que permanecen escondidos mientras no son solicitados. De este modo, el texto entero de las tres preguntas anteriores podría ser utilizado como etiquetas de cada una de las variables y, en caso que el nombre de la variable no fuera comprensible para el investigador, podría ser consultado inmediatamente. También, muchos análisis estadísticos que no tengan limitaciones de espacio podrían utilizar esas etiquetas de variables para acompañar los resultados producidos, facilitando de este modo su interpretación.

c) Posiciones de las variables. Las posiciones de las variables es un tercer aspecto que debería ser cuidado a la hora de codificar un archivo de datos. En general, la regla debería ser agrupar información relacionada (Davidson, 1996). Por ejemplo, las columnas podrían seguir el orden en el que se encuentran en un cuestionario, o, cuando no haya un referente físico, se podría buscar una relación temporal (notas de una asignatura en la primera evaluación, en la segunda, etc.) o lógica (contenidos similares, etc.). Estas posiciones simplemente no deberían dejarse indeterminadas y debería tenerse cuidado en que los programas utilizados no realicen manipulaciones sobre este orden utilizando criterios impropios (SPSS 6.0 por ejemplo ordenará las variables por orden

alfabético lo cual puede dar lugar incluso a resultados incorrectos en ciertos cálculos -p.e escalamiento multidimensional-). Otros programas de gestión de datos permiten crear diferentes ordenaciones de las variables lo cual puede ser útil en ciertas ocasiones.

- **Valores**

La codificación de los valores que se introducirán en la tabla está limitada por los tipos de datos normalmente admitidos en paquetes estadísticos. Para ellos generalmente sólo existen dos tipos de datos:

- **Datos de tipo texto:** Son textos de longitud variable, estando en ocasiones limitado el número máximo de caracteres admisible (un número usual para este límite es el 8). A menudo reciben el nombre de variables alfanuméricas. Asimismo, a menudo no pueden utilizarse ciertos caracteres tales como símbolos utilizados para operaciones matemáticas (+, -, \*, /) u otros especiales (@, •, y, %). Cuando existe un límite máximo para los datos de tipo texto es posible que exista el tipo texto largo, no utilizable en cálculos matemáticos, pero interesante para añadir comentarios o explicaciones a los datos. Siempre que se utilicen este tipo de datos debería tenerse cuidado en averiguar si el programa utilizado distingue entre una palabra escrita entre mayúsculas y minúsculas o no (p.e. si EDUCACION es igual a educación para el programa).

- **Datos de tipo numérico:** En ocasiones los paquetes distinguen entre datos numéricos enteros y datos numéricos con decimales. Los primeros ocupan menos espacio en disco y los cálculos con ellos pueden ser más rápidos en ocasiones.

Vemos los siguientes tipos de datos usando la clasificación de Stevens y como los codificaríamos en un paquete estadístico.

*Variables nominales.*

Para representar una variable nominal los paquetes estadísticos ofrecen como opción más adecuada, aparentemente, las variables de tipo texto. No obstante, esta no suele ser la opción tomada debido a que ello alargaría el proceso de introducción de datos innecesariamente. Por ejemplo, si tuviéramos una variable nominal "religión" podríamos introducirla de las tres maneras mostradas en la tabla 2.11:

<i>Texto completo</i>	<i>Abreviaturas</i>	<i>Código num.</i>
<b>Cristiano</b>	<b>C</b>	<b>1</b>
<b>Cristiano</b>	<b>C</b>	<b>1</b>
<b>Musulmán</b>	<b>M</b>	<b>2</b>
<b>Musulmán</b>	<b>M</b>	<b>2</b>
<b>Budista</b>	<b>B</b>	<b>3</b>

Tabla 2.11. Codificación de variables nominales.

Las tres columnas ofrecen la misma información. No obstante, la primera de las columnas es excesivamente costosa de introducir y seguramente daría lugar a un gran número de errores por lo que no resultaría aconsejable. La segunda es más adecuada pero presenta el inconveniente de obligar al sujeto que introduce los datos a mover las manos desde la parte numérica del teclado a la parte alfabética lo cual enlentecería el proceso y quizás daría lugar a errores. El tercer método no obstante no está libre de problemas. Debido a que los números no tienen significado, cuando el número de categorías es muy grande o hay muchas variables codificadas de este modo puede resultar difícil mantener en la memoria cada uno de los valores. Además, las equivocaciones son más difíciles de corregir.

En cualquier caso, independientemente de como introduzcamos los datos, más tarde nos gustaría ver nuestros resultados etiquetados con los textos de las categorías completas (la primera columna) en lugar de las otras dos. De este modo, los análisis posteriores aparecen con más información en pantalla y resulta más simple seguirlos. Para ello, los paquetes estadísticos ofrecen un comando de re-codificación (véase en secciones posteriores) que permitiría definir una nueva variable o modificar una existente de modo que, por ejemplo, el valor "B" sea transformado a "Budista" para todos los casos.

Otra opción es utilizar *etiquetas de valores*, una idea que resulta similar a las etiquetas de variables comentadas anteriormente. Estas etiquetas de valores corresponderían a los textos de mayor longitud asociados a cada uno de los valores posibles que puede tomar una variable. Estas etiquetas permanecerían escondidas hasta el momento que se deseara consultarlas.

Un caso especial de variables nominales son las variables dicotómicas. Estas son usualmente representadas por los valores 0 y 1, aunque esta no es una regla estricta.

Utilizar estos valores presenta la ventaja de que ciertos análisis estadísticos (p.e análisis de regresión) pueden ser especificados sin realizar transformaciones a las variables.

*Variables ordinales.*

Las variables ordinales que representan rangos o jerarquías necesitan en cambio variables numéricas con números decimales. No obstante, este último tipo de variables muy a menudo no son introducidas directamente sino que son producto de una transformación de datos pertenecientes a otras variables. Por ejemplo, supongamos que tenemos datos acerca de los presupuestos de equipos deportivos y sus resultados al final de una temporada (v. tabla 2.12). La variable partidos ganados es la que tiene interés en este caso. Esta variable es de tipo cuantitativo e indica el número de partidos ganados por cada uno de los equipos. No obstante, debido a motivos teóricos nos podríamos plantear si esta variable debería ser manejada así o bien debería considerarse únicamente atendiendo a su posición jerárquica. La cuarta variable es el resultado de un comando de un paquete estadístico que lleva a cabo justo esto. Algunos procedimientos estadísticos (p.e. correlación de Spearman) llevan a cabo esta transformación internamente.

Equipos	Presupuesto	Partidos ganados	Rangos de Partidos ganados
A	100000000	20	7
B	80000000	18	5.5
C	80000000	18	5.5
D	80000000	15	4
E	40000000	14	3
F	10000000	13	1.5
G	10000000	13	1.5

Tabla 2.12. Transformación en Rangos o Jerarquías

*Variables de intervalo/razón.*

Las variables cuantitativas (intervalo y razón) deberían introducirse como variables numéricas en el ordenador. Es conveniente distinguir entre los casos discretos y continuos para de ese modo trabajar con archivos de tamaño inferior, aunque este criterio hoy en día ha perdido la importancia que tuvo en el pasado.

Las puntuaciones deberían introducirse con el mayor nivel de precisión posible. No resulta adecuado perder nivel en cuanto a medición en nuestros datos simplemente por una decisión de codificación. Esto suele ocurrir cuando se opta por tomar una variable de

intervalo/razón como si estuviera compuesta de categorías que van de un valor a otro. Por ejemplo, en un estudio se podría optar por preguntar a los sujetos si su Edad está entre: a) 0-20 años; b) 21-30 años; c) 31-40 años; d) más de 40. La variable así recogida podría ser tratada como perteneciente a nivel categorial o, quizás, ordinal. Desde el punto de vista del análisis, preguntar la edad sería mucho más ventajoso.

<ul style="list-style-type: none"><li>• <i>ID: Identificador. Un número o etiqueta que identifica individualmente a cada sujeto.</i></li><li>• <i>GRUPO: 1= Método de enseñanza A; 2= Método de enseñanza B; 3= Método de enseñanza C; 4= Faltante.</i></li><li>• <i>SEXO: 0= Hombre; 1=Mujer; 3=Faltante.</i></li><li>• <i>MAT: Matemáticas. Puntuaciones entre 0 a 10 divididas por 2. -99=Faltante.</i></li><li>• <i>AC1Mat: ¿Te gusta hacer los deberes de Matemáticas? De 1 a 7. Valores altos indican actitud positiva. -9=Faltante.</i></li><li>• <i>AC2Mat: ¿Te gusta ir a clase de Matemáticas?. Valores altos indican actitud positiva. -9=Faltante.</i></li><li>• <i>AC3Mat: ¿Te resulta cansado ir a clase de Matemáticas? Valores altos indican actitud negativa. -9=Faltante.</i></li><li>• <i>E-EPQ: Puntuaciones directas en la escala de Extroversión del cuestionario EPQ. -9=Faltante.</i></li></ul>
---

*Tabla 2.13. Ejemplo de libro de códigos*

La unidad de medida en que están recogidos los datos debería tenerse en cuenta. En muchas ocasiones, uno puede cambiar la unidad (por ejemplo introducir miles de pesetas en lugar de pesetas) para disminuir el número de dígitos a introducir.

Las *fracciones de recuentos* (proporciones o porcentajes) son introducidas mejor, en muchas ocasiones, como los recuentos en su forma directa para más tarde realizar una transformación que produzca por ejemplo los porcentajes buscados. Otros tipos de datos pueden también ser introducidos mejor teniendo en cuenta que más tarde pueden ser transformados a un formato más adecuado.

### **2.5.1. El registro de la codificación**

A pesar de que los usuarios suelen pensar que resulta imposible olvidar los detalles acerca de unos datos y que el esfuerzo de transcribir los detalles resulta innecesario, esta



es una práctica que resulta muy aconsejable. En la tabla 2.13 se puede ver un ejemplo de lo que es denominado un *libro de códigos*:

Como vemos, en el libro se indican las categorías de cada pregunta, los códigos de valores faltantes (utilizando la convención de usar valores fuera del rango posible y además negativos pero véase los comentarios sobre codificación de valores faltantes más adelante), una pequeña descripción y el nombre abreviado de la variable. Se indica por ejemplo si una pregunta está en una dirección u otra (v. AC1Mat v. AC3Mat) y cualquier otra información relevante. En la tabla 2.14 hay un ejemplo del aspecto que podrían tener unos datos correspondiente al libro de la tabla 2.13.

ID	GRUPO	SEXO	MAT	AC1-MAT	AC2-MAT	AC3-MAT	E-EPQ
1	1	1	-99	1	-9	7	17
2	1	0	3	2	1	5	17
3	1	1	1	5	3	4	23
4	1	1	-99	7	6	1	23
5	2	0	1	3	7	3	21
6	2	0	4	2	4	2	21
7	2	0	1	1	3	1	23
8	3	1	1.5	3	2	1	-9
9	3	0	1	2	1	7	21

Tabla 2.14. Ejemplo de datos siguiendo el libro de códigos de la figura 2.13

Este libro puede ser aceptable para uso interno de un proyecto pequeño. El ICPSR (1997) indica que un grupo de trabajo interdisciplinario está desarrollando lo que denominan un SGML (Standard Generalized Markup Language- Lenguaje de Etiquetado Generalizado) para la construcción de libros de códigos en ciencias sociales, lo cual facilitaría enormemente la transmisión de datos entre diferentes grupos de investigación. Mientras tanto, indican que la siguiente información acerca de las *variables* (otras áreas no son repetidas aquí) debería ser proporcionada:

1. El texto exacto de la pregunta o el significado exacto del dato individual.
2. El número del ítem. P.e. 3a.
3. El universo de información. Es decir, a quién se le preguntó realmente. En otras palabras, si hay un patrón de salto tal que algunos ítems no fueron preguntados a todos los respondentes, la información debería estar en la misma página que el resto de la información para ese ítem.
4. Distribuciones de frecuencias o estadísticos resumen para el ítem. Estos datos deberían incluir información acerca de valores faltantes.

5. Códigos de datos faltantes.
6. Información acerca de asignación y comprobación de los datos. Si la contestación ha sido asignada para algunos sujetos o si se ha llevado a cabo una comprobación cuidadosa, esta información debería ser indicada. Este texto presenta más adelante secciones bastante amplias acerca de comprobación y asignación de datos .
7. Para variables construidas, se debería de dar información acerca de como se realizaron los cálculos. Si alguna de las variables utilizadas en los cálculos tenía valores faltantes, se debería indicar qué se realizó en este caso. El código exacto utilizado para construir la variable sería también interesante.
8. Significado exacto de los códigos. Para cada valor que puede adoptar una variable, el libro de códigos debería mostrar la interpretación de ese código. Para algunas variables, tal y como los códigos de ocupación, la información podría aparecer en los apéndices.
9. Localización en el archivo de datos. Indicar la columna en la que están los datos.

### **2.5.2. Casos especiales**

Algunos valores o variables necesitan una codificación especial. Serán tratados a continuación.

#### *a) Valores faltantes, ausentes o no-respuestas.*

En la vida real muy a menudo los datos no están completos. Los sujetos deciden no contestar las preguntas que les resultan incómodas, existen fallos en la información o resulta imposible obtener los valores buscados. Esta es una situación muy indeseable, ya que los análisis de datos pueden resultar viciados (los sujetos que contestan o dejan de contestar están sesgados hacia ciertas tendencias), insuficientes (muchos procedimientos estadísticos necesitan eliminar todos los datos de los sujetos que tienen algún dato faltante) o inestables (dependiendo del procedimiento estadístico utilizado, resultados que deberían ser iguales varían entre sí según el método utilizado para eliminar los valores faltantes).

Para Davidson (1996) los datos faltantes o similares suponen la mayor fuente de problemas en el procesamiento de datos siendo una codificación adecuada el primer paso para evitarlos.

Los datos faltantes pueden provenir de diferentes causas, las cuales deberían ser codificadas separadamente. El ICPSR cita las siguientes:

(i) Rechazo a responder.

(ii) No sabe. El sujeto fue incapaz de contestar, bien porque no tenía una opinión o porque la información pedida no estaba disponible.

(iii) Error de procesamiento. Por alguna razón no hay respuesta a la pregunta, aunque el sujeto proporcionó una en su momento. Esto puede deberse a un error del entrevistador, una codificación incorrecta, un fallo de la máquina y otros problemas.

(iv) No aplicable. El sujeto no fue preguntado acerca de esta cuestión por alguna razón. A menudo esto ocurre debido a "patrones de saltos" que ocurren, por ejemplo, cuando los sujetos que están desempleados no se les pregunta por el tipo de trabajo que están realizando. Otros ejemplos son conjunto de ítems que sólo se preguntan a submuestras aleatorias u otros preguntados a un miembro de la unidad familiar pero no a otros.

(v) No ajuste. Esto puede producirse cuando hay información que está siendo extraída de fuentes ya registradas y parte de ella no está disponible para todos los sujetos.

Podemos distinguir dos tipos de valores faltantes. El primero de ellos correspondería a uno genérico, que utilizaríamos cuando no deseamos hacer una distinción entre diferentes tipos de valores faltantes. Esto es denominado por paquetes estadísticos como el SPSS o SAS *valores faltantes del sistema*. Dependiendo del paquete estadístico que utilicemos esto podría indicarse mediante un punto (.), un espacio en blanco ( ) o un punto ancho (•). A menudo, un lugar vacío donde debería haber un valor será interpretado por los paquetes estadísticos como un valor faltante y el símbolo correspondiente será asignado automáticamente. El segundo tipo correspondería a lo que es denominado *valores faltantes del usuario*, los cuales proporcionan algún tipo de información que el paquete estadístico puede ser instruido para utilizar o no según nuestra conveniencia.

Conocer los símbolos y la conducta por defecto que utilice nuestro sistema de análisis de datos debería ser una de las prioridades en cuanto a su aprendizaje.

El ICPSR recomienda evitar usar espacios en blanco o huecos donde haya valores faltantes, aconsejando indicar explícitamente su existencia. Un error bastante grave (pero común) es asignar ceros a los valores faltantes. Algunos programas no específicamente diseñados para tratar con datos estadísticos pueden introducir ceros cuando se deja en blanco el espacio correspondiente a un dato.

Cuando deseamos indicar las causas de los valores faltantes utilizando una lista similar a la anterior los paquetes estadísticos pueden permitir utilizar *valores faltantes definidos por el usuario*. Estos valores son definidos en cada caso. Por ejemplo, un comando del tipo `<SI var1= -99 ENTONCES faltante_tipo_1>` indicaría que aquellos valores dentro de la variable 1 iguales a -99 serán considerados como faltantes del tipo 1.

La ventaja de realizar esta definición consiste en que los paquetes estadísticos permitirán ciertas operaciones que tienen sentido para los valores faltantes (por ejemplo, hacer recuentos del número de valores faltantes de cada tipo) y, automáticamente, los eliminarán para aquellas situaciones en que esto resulte apropiado. Además, muchos análisis permitirán opciones especiales apropiadas en cada caso.

El ICPSR recomienda hacer corresponder el tipo de los códigos de datos faltantes con el contenido del campo. Si el campo es numérico, los códigos deberían ser numéricos y si el campo es de texto, los códigos deberían ser de texto. Muchos investigadores utilizan valores numéricos que están muy por encima o muy por debajo de los que la variable puede tomar legalmente para así distinguirlos más fácilmente. Esto no presenta problemas siempre y cuando se defina correctamente el valor como faltante.

*b) Valores imposibles.*

En ciertos casos, los paquetes estadísticos realizarán codificaciones de datos *por sí mismos* (Jaffe, 1994). Un caso muy habitual es cuando al realizar una operación sobre los datos el resultado obtenido es imposible para ciertos casos. Por ejemplo, si dividimos una variable por otra y en la segunda algunos valores son iguales a cero el resultado será para algunos casos imposible. En este caso, el sistema adjudicará por ejemplo un valor de faltante.

Otro caso que producirá resultados imposibles es cuando se realice una operación sobre un caso faltante. Por ejemplo, a un grupo de sujetos se les suma las notas en dos evaluaciones y luego se divide por dos para obtener el promedio, pero algunos de ellos

no realizaron el segundo examen. Para estos el valor del promedio será *valorfaltante*. Cualquier otro valor significaría una asignación de valor que debería ser justificada cuidadosamente (dividir por dos la nota de la primera evaluación por ejemplo implicaría asignar cero a la segunda evaluación lo cual puede no ser muy apropiado).

*c) Variables de respuesta múltiple*

En otro lugar hemos descrito cómo las variables abiertas pueden producir un tipo de variables que hemos denominado de respuesta múltiple. Este tipo de preguntas pueden aparecer en una variedad de situaciones. Podemos definir las preguntas de respuesta múltiple como aquellas en las que los entrevistados pueden dar un número indeterminado de respuestas. Por ejemplo, a la pregunta de la figura 2.15 un entrevistado podría dar entre cero a catorce respuestas diferentes (y si permitiéramos que la alternativa "otros" indicara explícitamente cuáles, todavía más).

**¿Qué deportes le gustan?**

<input type="checkbox"/> Auto/Moto	<input type="checkbox"/> Ski
<input type="checkbox"/> Atletismo	<input type="checkbox"/> Montaña
<input type="checkbox"/> Caza	<input type="checkbox"/> Deportes Náuticos
<input type="checkbox"/> Ciclismo	<input type="checkbox"/> Tenis
<input type="checkbox"/> Hípica	<input type="checkbox"/> Gimnasia
<input type="checkbox"/> Fútbol	<input type="checkbox"/> Otros
<input type="checkbox"/> Golf	
<input type="checkbox"/> Pesca	

*Figura 2.15: Pregunta con respuestas múltiples*

Existen dos maneras de codificar esta información (Norusis, 1988) utilizando el formato conocido de columnas y filas y cuando los describamos veremos las dificultades que plantean este tipo de variables.

El primero de ellos se denomina de *dicotomías múltiples* e implica asignar a cada una de las alternativas de la pregunta una variable. Nuestra tabla de datos para seis sujetos podría ser la siguiente. En ella un uno indica que el sujeto dió esa alternativa como respuesta. Un cero indica que no la señaló.

ID	Aut/	Atlet	Caza	Cicli	Hípi	Futb	Golf	Pesc	Ski	Mon	Náuti	Teni	Gim	Otro
1	1	0	0	0	0	1	0	0	0	0	0	0	1	0
2	0	1	0	0	0	1	0	0	1	0	0	0	1	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Tabla 2.16. Tabla de datos utilizando dicotomías múltiples

El segundo tipo de codificación se denomina de respuestas múltiples e implicaría examinar el número máximo de respuestas que haya dado algún sujeto y crear un número de variables igual a éste. A continuación se asignaría un código a cada una de las alternativas de respuesta y se introduciría en las columnas correspondientes. En nuestro caso, el número de columnas necesitadas es cuatro y la tabla de datos podría ser mostrada en la tabla 2.17 (las actividades deportivas han sido numeradas de 1 a 14 siguiendo el orden de la tabla 2.16).

ID	Var1	Var2	Var3	Var4
1	1	6	13	0
2	2	6	9	13
3	0	0	0	0
4	1	0	0	0
5	7	0	0	0
6	1	2	0	0

Tabla 2.17. Tabla de datos utilizando respuestas múltiples

Cada método tiene ventajas e inconvenientes. El primero necesita más columnas pero es probablemente más claro de visualizar. El segundo necesita menos espacio pero el número de columnas necesarias puede ser difícil de precisar. Además, si algunos sujetos dan la lista entera de posibles contestaciones como respuesta, el número de columnas necesario es el mismo en ambos casos.

En ambos casos se debería tener en cuenta que los valores asignados a cero se deberían contar como valores faltantes o de no respuesta.

Dependiendo del análisis estadístico a realizar los datos deberán estar organizados de una manera u otra. De modo tentativo, se podría decir que utilizar dicotomías múltiples

es probablemente el método más flexible a pesar que puede multiplicar enormemente el número de columnas necesarias.

Ambas comparten las dificultades para realizar operaciones en apariencia sencillas. Por ejemplo, calcular cuanta gente demuestra un interés por algún deporte (no uno en concreto, sino alguno de ellos) no es directo (habría que hacer un cálculo a través de cada fila para determinar si hay alguna contestación positiva, y luego hacer la suma de ese cálculo). También, en el caso de las respuestas múltiples, saber cuanta gente está interesada en un determinado deporte (digamos fútbol) no es sencillo si tenemos en cuenta que los paquetes estadísticos suelen tener problemas para hacer cálculos a través de varias variables. Debido a estas dificultades algunos paquetes estadísticos (p.e. SPSS) incorporan funciones específicamente diseñadas para tratar con este tipo de datos y que facilitan estas operaciones.

*d) Codificación para el análisis de variables categóricas.*

Ciertas técnicas estadísticas necesitan de una codificación previa de las variables categóricas para poder utilizarlas apropiadamente. Un ejemplo muy común es el análisis de regresión (Pedhazur, 1982) para así hacer frente a problemas habitualmente tratados desde el punto de vista del análisis de varianza, pero también surge en otros contextos (López, et al., 1997).

De modo esquemático, comentaremos el esquema de codificación mediante variables ficticias (dummy) y de efectos fijos.

La codificación ficticia consiste en convertir una variable categórica en  $k-1$  variables que indican mediante unos si un sujeto dado está en una variable dada y ceros que no. En la figura 2.18 podemos ver un ejemplo. La variable Religión con tres categorías (M, C y B) ha sido codificada en dos variables M y C. Aquellos sujetos que tenían como religión C tienen un 1 en la variable correspondiente y 0 en la otra. La tercera categoría, B, sin embargo, no tiene una variable propia y corresponde a los sujetos que tienen simultáneamente un cero en las dos variables M y C.

ID	Religión	M	C
1	C	0	1
2	M	1	0
3	C	0	1
4	M	1	0
5	B	0	0
6	B	0	0
7	C	0	1

Figura 2.18. Codificación ficticia de una variable

La codificación mediante efectos fijos es igual a la anterior pero la categoría que no tiene variable propia recibe -1 en las variables en lugar de 0. Corresponde a la figura 2.19.

ID	Religión	M	C
1	C	0	1
2	M	1	0
3	C	0	1
4	M	1	0
5	B	-1	-1
6	B	-1	-1
7	C	0	1

Figura 2.19. Codificación por efectos fijos de una variable

La justificación de este procedimiento, así como la explicación de un tercer tipo de codificación (ortogonal) escapan a los objetivos de este trabajo. Pedhazur (1982) proporciona una buena explicación de los tres tipos de codificación así como de las consecuencias para la interpretación de los análisis en Regresión Múltiple.

*e) Variables ponderadoras.*

Una variable ponderadora multiplica los valores en otras variables para de este modo tener en cuenta el número de elementos que incluye cada fila. Por medio de estas variables por tanto podemos ahorrarnos el introducir datos de índole repetitiva, al poder simplemente teclear el valor a considerar y el número de veces que se repite. Los datos de la tabla 2.20 corresponden a dos variables nominales en el formato habitual de filas por columnas. Sólo se muestran los 6 primeros datos de cada variable.



<i>Sujeto</i>	<i>Género</i> (H, M)	<i>Color</i> (R, V, A)
1	M	R
2	H	V
3	M	A
4	M	A
5	H	V
6	M	R

Tabla 2.20. Dos variables nominales codificadas como variables separadas

No obstante, es relativamente fácil encontrar esa misma información utilizando una tabla como la 2.21.

<b>Color</b>	<b>Género</b>	
	<i>Hombres</i>	<i>Mujeres</i>
<i>Verde</i>	250	300
<i>Rojo</i>	589	265
<i>Amarillo</i>	235	892

Tabla 2.21: Datos en la tabla 2.19 en forma de tabla.

En esta tabla se refleja el recuento de las correspondencias entre categorías de la tabla mostrada en la tabla 2.20 (pero en lugar de sólo 6 casos aquí se mostrarían un total de 2531 casos!). Obviamente, cuando son apropiadas estas tablas son muy convenientes al ocupar mucho menos espacio.

A menudo los paquetes estadísticos nos permitirán introducir la información siguiendo el formato de la tabla 2.22.

<i>Sexo</i>	<i>Color</i>	<i>Variable ponderadora</i>
H	R	589
H	V	250
H	Am	235
M	R	265
M	V	300
M	Am	892

Tabla 2.22. Datos en la tabla 2.21 en forma de columnas.

Posteriormente sería necesario utilizar un comando que indicara al paquete estadístico que en la variable ponderadora se indica el número de valores que corresponde a cada caso.

Este sistema también puede utilizarse cuando se trata de variables numéricas. Por ejemplo, los sueldos en una empresa podrían estar como en la tabla 2.23.

<i>Salario</i>	<i>Variable ponderadora</i>
100000	120
150000	50
250000	10
500000	1

Tabla 2.23 Codificación de una variable numérica mediante otra ponderadora.

En este caso, la variable ponderadora determina cuantos sujetos cobran el sueldo señalado a la izquierda. Si quisiéramos obtener la media de los salarios en esa empresa podríamos introducir los datos de ese modo, ponderar la primera variable por la segunda y solicitar la media (ponderada). El resultado sería 124309.39 (y no 250000). Tener en cuenta que el número de casos es 181 (no 4).

El inconveniente de utilizar variables ponderadoras es que la información correspondiente a los individuos desaparece. De este modo, esta codificación implica *agregación* de datos (un concepto que es discutido en el apartado de transformaciones).

*f) Variables selectoras y de grupo.*

Los paquetes estadísticos suelen utilizar los términos de variable selectoras y de grupo. Ambas pueden ser entendidas como variables categóricas pero que son utilizadas

de una manera concreta por estos programas. Las variables selectoras son variables que indican si un valor es seleccionado para ser utilizado en un análisis u operación (es decir, son una variable binaria). Las variables de grupo son variables que instruyen al programa para realizar la operación u análisis *para cada una de las categorías existentes en ella*. Por ejemplo, si tenemos notas escolares y sólo queremos utilizar alumnos que superen cierta calificación para calcular unos estadísticos descriptivos (medias, desviaciones, etc.) podemos construir una variable selectora que indique si un alumno ha superado o no esa nota. En cambio, si lo que queremos es obtener esos estadísticos descriptivos para los alumnos *en cada colegio* podemos utilizar una variable de grupo que indique la pertenencia de cada estudiante a cada colegio. El resultado será una media, una desviación, etc., para cada uno de los colegios.



# ***Creando Archivos de datos Computerizados: Introducción de datos***

## **3.1. Introducción**

En la situación actual, en la que los análisis se realizan principalmente utilizando computadores la tarea más pesada en un estudio de índole estadística puede ser la de introducir los datos de la manera correcta. Podemos considerar varios casos dependiendo de los medios de que dispongamos. En un extremo tenemos el de grupos de datos de tamaño pequeño que queremos introducir rápidamente nosotros mismos. En el otro se encontrarían proyectos de gran tamaño, que necesitan muchos individuos recogiendo y registrando datos en caso de realizarse esta tarea manualmente, que deben estar coordinados entre sí para realizar la labor de la manera más eficiente y libre de errores posible. Podemos plantear los siguientes escenarios:

- La información es registrada e introducida manualmente: Probablemente la forma más común de introducir los datos aunque existen varios grados de complejidad posible dentro de ella.

- La información es registrada en papel pero posteriormente es introducida de modo automático. Esta sería la situación en la que se utiliza un *scanner*, un lector de marcas o un aparato similar.

- La información es registrada directamente en el ordenador. En ella el usuario puede autoadministrarse las preguntas, o, también, el entrevistador puede introducir las respuestas directamente en el ordenador.

Estas posibilidades serán descritas a continuación.

### **3.2. La información es registrada e introducida manualmente**

Quizás la forma más habitual de introducir información en el ordenador sea utilizando el teclado (v. figura 3.1) a partir de un registro escrito tal y como un cuestionario o similar. Este método resulta adecuado para proyectos de tamaño reducido o medio en el que otros métodos más complicados pueden no resultar ventajosos en cuanto a costo. En este caso podemos distinguir dos grupos de programas que podrían ser utilizados para este propósito y que serán descritos a continuación. Estos serían los programas no diseñados específicamente para este propósito y aquellos que incorporan herramientas específicas para la introducción de datos estadísticos. En el primero de los grupos tenemos programas como editores u hojas de cálculo, los cuales fueron diseñados para llevar a cabo tareas respectivamente relacionadas con manejo de textos y de números pero que pueden ser utilizados para introducir datos de tamaño pequeño o medio. En el segundo de los grupos tenemos programas que funcionan en coordinación con paquetes estadísticos o que forman parte de sistemas generales de gestión de datos que permiten la creación de pantallas específicamente diseñadas para nuestros datos y que resultan especialmente útiles cuando aquellos que se van a encargar de la introducción no están acostumbrados a este tipo de tareas.



*Figura 3.1. Teclado de ordenador*

### 3.2.1. Editores y/o procesadores de texto

Quizás pueda parecer sorprendente, pero uno de los programas más útiles para la gestión de conjuntos de datos no excesivamente grandes (digamos de unos cuantos cientos de casos aunque el progresivo aumento de la capacidad de los ordenadores está aumentando sin duda este límite) son programas en principio diseñados para manejar texto y no números. Esto se debe fundamentalmente a la existencia del código ASCII (*American Standard Code for InterChange of Information*) y a su utilización como método para el intercambio de información de estructura simple a partir de programas de muy diversa índole. Aunque siempre sujeto a cuestiones dependientes de cada caso concreto, el código ASCII (o como suele ser denominado en los programas el formato Sólo Texto) permite pasar información entre hojas de cálculo, bases de datos, paquetes estadísticos y otros programas debido a que prácticamente todos ellos han sido diseñados para aceptarlo y utilizarlo. Este código, debido a su simplicidad, necesita, no obstante, utilizar ciertas convenciones que serán descritas con más detalle en el apartado de importación y transferencia de datos, pero, por el momento mostraremos la siguiente imagen de como se ve un archivo de tipo estadístico en un procesador de textos (figura 3.2).

VALENCIANO	LENGUA	IDIOMA	ES	MATEM	C/N
3	2	5	4	2	5
2	2	4	3	2	5
1	1	5	1	2	5
2	2	4	2	2	5
3	4	5	4	3	5
4	4	4	2	5	4
1	1	5	0	1	5
2	4	5	1	2	5
3	1	5	2	3	5
2	1	5	2	2	5
1	1	5	0	1	5
1	3	5	2	2	5
2	1	5	4	2	5
1	1	5	0	2	5

Figura 3.2. Archivo de datos estadísticos en formato ASCII en un procesador de textos

En este archivo se muestran caracteres que normalmente serían invisibles al usuario tal y como ¶ que indicaría cambio de párrafo o entre líneas y \* que serviría para señalar la separación entre columnas. La primera fila en este caso, tiene los nombres de las

columnas de datos. Una descripción más completa de este formato y otros relacionados se encuentra en el apartado de importaciones y transferencias de archivos.

La utilización de archivos de tipo texto permite realizar cambios en los datos utilizando la función de *Buscar y Reemplazar* de los editores de texto. Por ejemplo, un cambio muy habitual (en este lado del mundo) es sustituir las comas (,) por puntos (.). Las primeras indican valor decimal a la derecha en la mayoría de los países europeos salvo Gran Bretaña. Los segundos indican valor decimal en los países anglosajones. Esta convención se sigue habitualmente en los sistemas operativos y algunos programas de uso común (por ejemplo, hojas de cálculo) pero no en otros (algunos paquetes estadísticos). La función *Buscar y Reemplazar* puede ser utilizada tal y como se muestra en la figura siguiente para corregir esta situación.

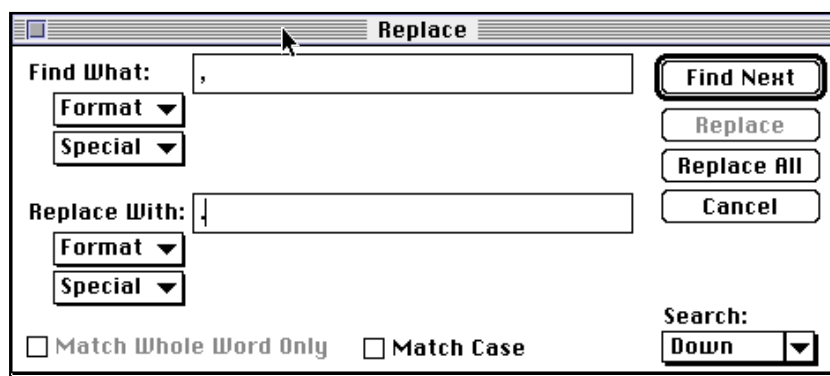


Figura 3.3. Función de reemplazar en un procesador de textos. En la parte superior se indica lo que se desea reemplazar y en la parte inferior aquello que lo reemplazará

Este tipo de cambios resulta relativamente primitivo y puede dar lugar a consecuencias totalmente imprevisibles. Sin embargo, dependiendo del programa utilizado las posibilidades son enormes y muy parecidas a las posibles en un paquete estadístico (cambiar caracteres especiales, buscar en varios archivos diferentes, distinguir entre mayúsculas o minúsculas, etc.). En el pasado, resultaba habitual utilizar el lenguaje Basic para realizar este tipo de transformaciones debido a su simplicidad, adecuada para programas cortos y de uso interno. En la actualidad, debido a las capacidades que poseen los programas de uso general es posible realizar tareas equivalentes utilizando la función de *Buscar y Reemplazar*.



### 3.2.2. Hojas de cálculo

Aunque los editores de texto pueden ser utilizados para introducir los datos directamente, y en algunos casos pueden ser incluso más recomendables que alternativas más sofisticadas (por ejemplo, cuando el ordenador a utilizar sea obsoleto) éstos presentan inconvenientes relativamente obvios. Por ejemplo, cuando se trabaja con muchas variables, las columnas no están delimitadas con claridad y resulta fácil saltarse alguna cuando se introducen datos. Esto está mejorado en los programas denominados hojas de cálculo.

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>
<b>1</b>	VALENCIANO	LENGUA	IDIOMA	<b>HISTORIA</b>	MATEMAT	CIENCIAS
<b>2</b>	3	2.5	4	<b>2.5</b>	2	3
<b>3</b>	2	2	3	<b>2.5</b>	3	2.5
<b>4</b>	1	1.5	1.5	<b>2</b>	1	1
<b>5</b>	2	2.5	2.5	<b>2.5</b>	1	1.5
<b>6</b>	3.5	4	4	<b>3</b>	1	3.5
<b>7</b>	4	4	4	<b>2.5</b>	4	4
<b>8</b>	1.5	1.5	0.5	<b>1.5</b>	1	1.5
<b>9</b>	2	4	1.5	<b>2.5</b>	1.5	4
<b>10</b>	3	1.5	2	<b>3.5</b>	1	2.5
<b>11</b>	2	1.5	2	<b>2</b>	1	1
<b>12</b>	1	1	0.5	<b>1.5</b>	1	1
<b>13</b>	1	3	2	<b>2</b>	1.5	3
<b>14</b>	2.5	1	4	<b>2</b>	1	1.5
<b>15</b>	1.5	1	0.5	<b>2</b>	1	1
<b>16</b>	2	2	3	<b>2</b>	1	1.5
<b>17</b>	1	1.5	2	<b>1.5</b>	1	1.5

Figura 3.4. Datos en hoja de cálculo

Una hoja de cálculo presenta ventajas obvias respecto a los editores de texto. Por ejemplo, cuando el número de variables/columnas es grande, los datos vienen como filas/casos y la anchura de la hoja de cálculo supera el tamaño de pantalla disponible, se produce una situación incómoda. Por otro lado, resulta fácil establecer ciertas marcas en las columnas (por ejemplo, usando el tipo de letra **negrita** como en la columna etiquetada como Historia) para realizar comprobaciones a la hora de introducir los datos. También, muchas de ellas permiten mantener visible la primera fila (la dedicada a los nombres de las variables) durante el tiempo que dura la introducción de datos, cosa que en un editor de textos resulta más complicado.

Las hojas de cálculo son suficientes para la introducción de conjuntos de datos de tamaño moderado en los que la utilización de métodos más sofisticados no resultaría apropiado. En caso de requerirse una introducción más sofisticada, que implicara gran cantidad de datos, con una estructura más compleja que la de una tabla y llevada a cabo

por un grupo de gente con experiencia media o baja con ordenadores sería conveniente la utilización de programas específicamente diseñados para tal fin.

Una alternativa interesante en ocasiones son las tablas que los propios paquetes estadísticos incorporan. Hay que advertir, sin embargo, que los programas de hoja de cálculo resultan a menudo mucho más sofisticados y flexibles que estos programas, por lo que tradicionalmente han sido utilizadas preferentemente a aquellas. Las últimas versiones de los programas estadísticos, sin embargo, han mejorado bastante este aspecto así que es posible que pasen a convertirse en una solución con mayor aceptación en el futuro próximo (si no lo han sido ya).

### 3.2.4. Bases de datos

Las bases de datos son uno de los tipos de programas que hemos denominado basados en formularios. Estos permiten el diseño de lo que se conoce como presentaciones que pueden ajustarse para facilitar la introducción de datos. Además, estos programas suelen incorporar la posibilidad de trabajar con estructuras más complejas de los datos, además de permitir validaciones de la información introducida. El otro tipo son programas específicamente orientados a la introducción de datos estadísticos y que suelen ser distribuidos por empresas especializadas en programas de este tipo.

Las bases de datos tienen como finalidad fundamental el almacenamiento y recuperación de todo tipo de información, no estando dirigidos de una manera específica a la de tipo estadístico.

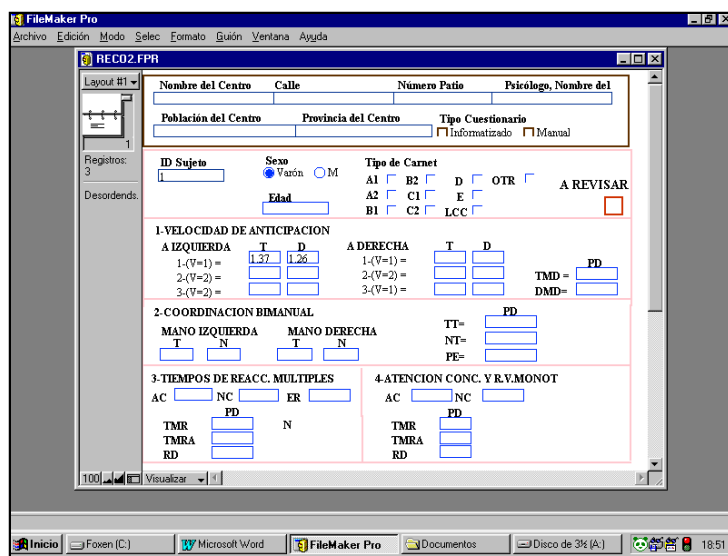


Figura 3.5. Plantilla creada utilizando un programa de bases de datos

En la figura 3.5 se muestra una pantalla de FileMaker dirigida a la introducción de unos datos correspondientes a una prueba de evaluación de conductores. Este formato corresponde muy cercanamente al del cuestionario que sirvió para recoger la información puesto que se consideró que el sujeto que llevaría a cabo la introducción vería facilitada la tarea al potenciarse esa semejanza. Además, se incluyeron elementos gráficos que servirían para contestar ciertas opciones. Esta plantilla puede utilizarse sin despegar las manos del teclado, evitando los inconvenientes del uso excesivo del ratón.

Entre las opciones disponibles para facilitar la introducción de datos en estos programas resulta interesante comentar las siguientes:

- Introducir automáticamente para cada nuevo registro información acerca del día, la hora, el autor (en caso que sea necesario identificarse para utilizar el programa) en que fue creado. También lo mismo acerca de modificaciones realizadas.

- Introducir automáticamente un número de serie para de ese modo identificar los registros de la base de datos.

- El valor del registro anterior. Para facilitar la introducción de datos que se repiten en cierto grado.

- El valor de un cálculo realizado sobre valores introducidos en otros campos. De este modo es posible realizar transformaciones sobre los datos de modo continuo pudiéndose también realizar ciertas comprobaciones.

- Valor obtenido de una tabla relacionada con la actual. En este caso, la estructura de los datos resulta de tipo relacional y la información puede ser consultada en otra tabla a partir de la que ha sido introducida hasta el momento. Este concepto se explica con más detalle en la sección dedicada a estructura de las bases de datos.

- Prohibición de modificación del valor. De tal modo que el usuario no pueda cambiar ya sea por error u otra razón el dato introducido automáticamente mediante alguna de las otras opciones.

Otras opciones hacen referencia a la validación de las entradas de datos. Comentaremos las más sencillas dejando para un apartado específico posterior las más avanzadas.

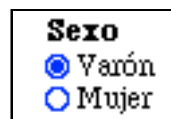
- Tipo de datos: Los tipos de datos más comunes son los numéricos o de texto. Por medio de esta opción podemos por ejemplo evitar que el usuario introduzca una letra en un campo que sólo admite valores numéricos. Otros datos más complicados son los de

fecha y hora. Una fecha debe obligatoriamente incluir día, mes y año, admitiéndose abreviaturas apropiadas. Si alguien desea introducir sólo el día y el mes pero no el año de un suceso deberá utilizar un formato diferente al de fecha con este propósito. El campo de hora puede o no incluir minutos y segundos.

- Dato no vacío: No se admite dejar en blanco un campo.
- Dato único: El dato no puede ser igual a cualquier otro previamente introducido.
- Dato previamente existente. El dato tiene que ser igual a cualquier otro previamente introducido.
- Miembro de un diccionario de valores. El valor no puede ser diferente de una lista de datos predeterminada.

Existen otras opciones que podríamos denominar gráficas que facilitan la introducción de datos y que pueden considerarse prácticamente como standard en relación con los ordenadores actuales:

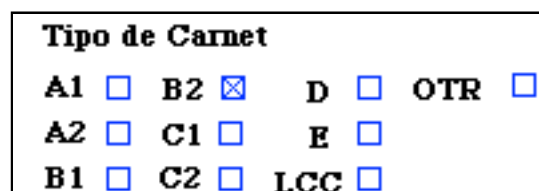
- Uso de botones de radio para indicar alternativas excluyentes. Este es el caso de la figura 3.6:



Sexo  
 Varón  
 Mujer

*Figura 3.6. Botones de radio para introducir una variable*

- Uso de cuadros marcados con cruces para indicar que varias alternativas son aceptables (figura 3.7).



Tipo de Carnet  
A1  B2  D  OTR   
A2  C1  E   
B1  C2  LCC

*Figura 3.7. Cajas para indicar el tipo de carnet*

- Cuadros para indicar los tamaños aproximados de los datos o simplemente para mejorar el aspecto gráfico.

- Menús: Adecuadas para las mismas situaciones que los botones de radio (sólo una opción es válida) pero aprovechan mejor el espacio de la pantalla ya que permanecen ocultas hasta que es necesario utilizarlas. En la figura 3.8 se han añadido dos opciones que resultan de interés: Permitir la introducción de un valor diferente a los existentes en el menú y editar la lista para permitir añadir nuevos valores a la propia lista.

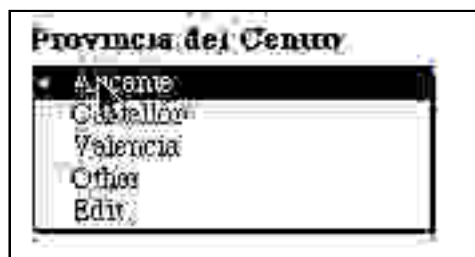


Figura 3.8. Lista desplegable para indicar la provincia

- Listas desplegables: Semejantes a los menús pero permiten utilizar el teclado para introducir los datos (figura 3.9).



Figura 3.9. Menú para indicar la provincia

Cuando la introducción de datos vaya a ser realizada por varios individuos bajo la supervisión de alguien encargado específicamente de la gestión de los datos y la creación de los formularios resulta habitual que surjan dudas entre ellos que deberían ser resueltas en otro momento. Para indicar esta situación en este ejemplo se incluyó lo siguiente:

- Un campo denominado A REVISAR.
- Un campo de comentarios. Donde se puede teclear una descripción rápida del problema.

- Alternativa de OTROS en el campo dedicado al carnet.

Los programas de bases de datos suelen incorporar también un lenguaje de programación que permite diseñar "caminos" a lo largo de un cuestionario. Por ejemplo, ciertas preguntas no serán aplicables a todos los sujetos o, dependiendo de lo contestado anteriormente, podemos querer explorar con más detalle ciertas áreas. La introducción de datos puede llegar a ser bastante complicada por lo que su automatización resulta de especial interés. Un ejemplo se encuentra en la figura 3.10. En él, la primera pregunta acerca de si ha pasado la Inspección Técnica de Vehículos anteriormente hace que las otras dos preguntas no sean aplicables y deban ser saltadas cuando la respuesta es negativa. Esta situación es de especial relevancia cuando el sistema utilizado para recoger las respuestas pasa por obtenerlas directamente de los entrevistados y estos saltos deben hacerse de modo automático.

<p><b>No Vale</b> <input type="checkbox"/></p> <p><b>Experiencia ITV</b></p> <p>1. ¿ITV anteriormente? <input type="checkbox"/></p> <p>2. ¿Última vez ITV? <input type="checkbox"/></p> <p>3. ¿Cuántas veces ITV? <input type="checkbox"/></p> <p>4. ¿Cómo supo de ésta ITV? <input type="checkbox"/></p>	<p><b>Pasos ITV</b></p> <p>10. ¿Conoce pasos ITV? <input type="checkbox"/></p> <p>11. ¿Pagar en banco? <input type="checkbox"/></p> <p>12. ¿Pagar en ITV? <input type="checkbox"/></p> <p>13. ¿Ir directamente? <input type="checkbox"/></p> <p>14. ¿Mecánico antes? <input type="checkbox"/></p> <p>15. ¿Sabía podía elegir día y hora? <input type="checkbox"/></p>	<p><b>Alternativas</b></p> <p>22. ¿Alternativa pago? <input type="checkbox"/></p> <p>22. bis ¿Efectivo/Tarjeta? <input type="checkbox"/></p> <p>23. ¿Pagar antes o desp? <input type="checkbox"/></p> <p>24. ¿Alternativas llamada? <input type="checkbox"/></p> <p>25. ¿Comentario aparte? <input type="checkbox"/></p>
<p><b>Opinión ITV</b></p> <p>5. ¿Opinión favorable ITV? <input type="checkbox"/></p> <p>6. ¿Obligación ITV? <input type="checkbox"/></p> <p>7. ¿Bien precio ITV? <input type="checkbox"/></p> <p>8. ¿Mejora su seguridad ITV? <input type="checkbox"/></p>	<p><b>Contacto telefónico</b></p> <p>16. ¿Dificultad contacto? <input type="checkbox"/></p> <p>16. bis ¿Qué problemas ITV? <input type="checkbox"/></p> <p>17. ¿Han ofrecido a su gusto día y hora? <input type="checkbox"/></p> <p>19. ¿Trato de la telefon? <input type="checkbox"/></p> <p>20. ¿Información adicional? <input type="checkbox"/></p> <p>20. bis ¿Qué información? <input type="checkbox"/></p> <p>21. ¿Motivo Elección ITV? <input type="checkbox"/></p>	<p><b>Datos Personales</b></p> <p>26. ¿Edad? <input type="checkbox"/></p> <p>27. ¿Sexo? <input type="checkbox"/></p> <p>28. ¿Profesión? <input type="checkbox"/></p>
		<p><b>Datos coche</b></p> <p>29. ¿Tipo de vehículo? <input type="checkbox"/></p> <p>30. ¿Propietario? <input type="checkbox"/></p>

Figura 3.10. Ejemplo de formulario con saltos

Por último, los programas de base de datos incorporan la posibilidad de trabajar en red para de ese modo tener varios usuarios introduciendo datos simultáneamente en el mismo archivo y evitar tener que mezclar los diferentes archivos con posterioridad.

### **3.2.5. Programas diseñados para paquetes estadísticos**

Los programas de bases de datos presentan el inconveniente para el usuario centrado en el diseño de cuestionarios y posterior análisis de ser excesivamente complejos desde el punto de vista de los requisitos estrictamente necesarios para esta tarea. Esta complejidad puede, no obstante, ser rentable ya que los programas de bases de datos resultan muy apropiados para llevar a cabo otras tareas diferentes a la de gestión de datos estadísticos. Sin embargo, si las bases de datos no van a ser utilizadas para nada más que la construcción de cuestionarios con una estructura de tabla con vistas a su posterior análisis los programas comentados en esta sección están mucho más centrados en este objetivo.

En líneas generales las capacidades que hacen a estos programas diferentes de las bases de datos son las siguientes:

- **Diseño de formularios:** Las opciones son muy parecidas a las contempladas en los programas de bases de datos aunque el vocabulario y las opciones gráficas están más dirigidas a la elaboración de cuestionarios.

- **Reglas de salto para ciertos valores.** Cuando un sujeto contesta de tal modo que otras preguntas no le son aplicables se produce un salto al lugar en el que corresponde continuar.

- **Llenado automático con ciertos valores:** Determinadas preguntas hacen obvias las respuestas a otras preguntas relacionadas. Por ejemplo, alguien con un contrato en prácticas trabaja en su empresa menos de un año. Los programas de introducción automática de datos pueden ser programados para llevar a cabo estos llenados automáticamente.

Estas dos últimas opciones son especialmente importantes cuando tratemos cuestionarios que son contestados por los propios entrevistados directamente.

- **Manejo de preguntas a las que es posible dar varias respuestas** (y su número puede ser difícil de anticipar de antemano). Por ejemplo, nombrar marcas de productos, o figuras del deporte conocidas por el entrevistado, etc.

- **Conexión con paquetes estadísticos.** Generando archivos de datos que pueden ser leídos directamente por éstos además de contemplar las convenciones que les son propios. Por ejemplo, programas que ofrecen conexión con SPSS ofrecen la posibilidad de crear etiquetas de valores y de variables compatibles con él.

• Procedimientos para comparar información introducida dos veces. Este método será explicado con más detalle en la sección acerca de depuración de datos.

A continuación describiremos algunos programas disponibles dentro de esta categoría.

Edwin soporta la creación de cuestionarios del siguiente modo. El usuario introduce cada una de las preguntas en una pantalla como la siguiente. En ella se escribe el texto de la pregunta, se selecciona el tipo de variable y si es posible contestar una o varias alternativas. El programa proporciona valores por defecto para las etiquetas de valores y de variables siempre que es posible (figura 3.11).

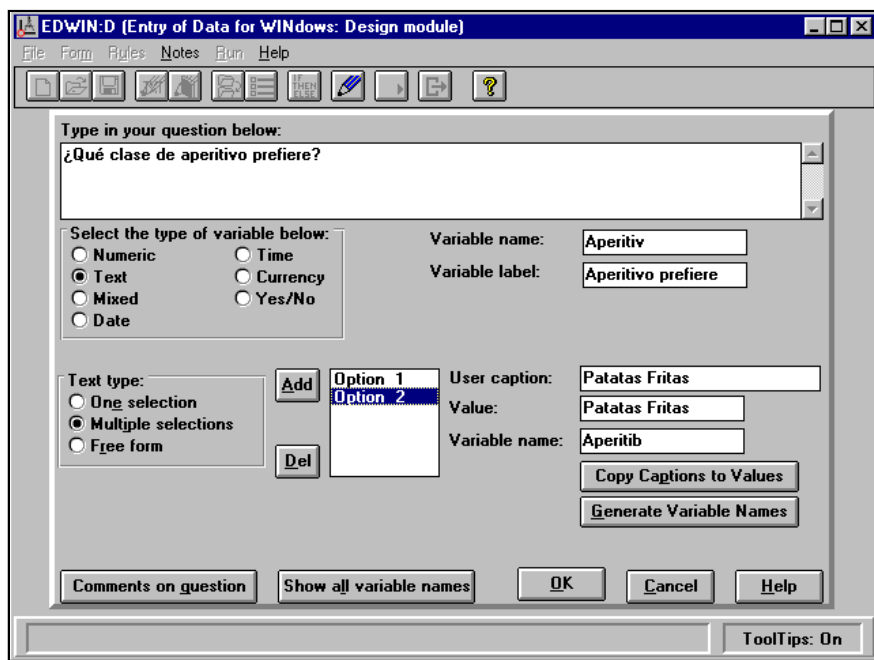


Figura 3.11. Definición de variables en Edwin

La lista de variables y las preguntas asociadas se mantienen en una lista como la siguiente (figura 3.12).



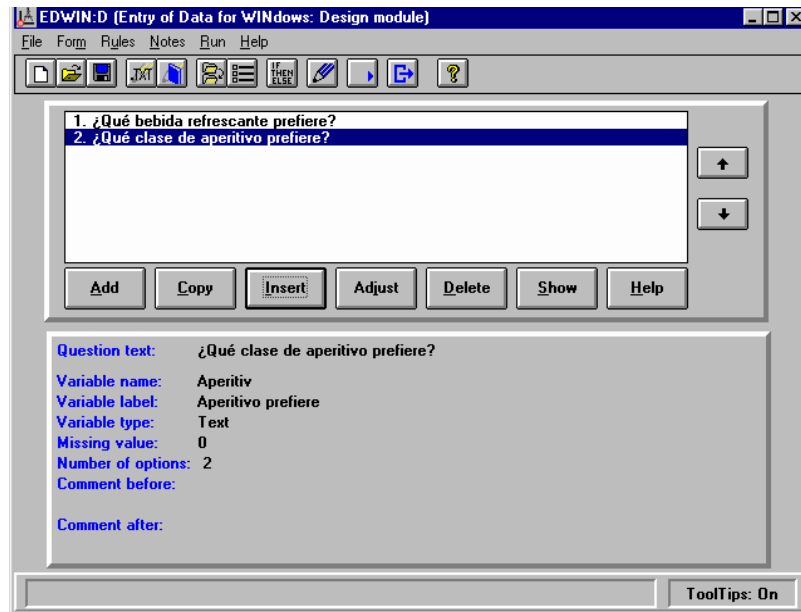


Figura 3.12. Lista de preguntas/variables en Edwin

Este programa permite la introducción de los datos para cada sujeto pregunta a pregunta. Además, exporta los datos a una gran variedad de formatos de paquetes estadísticos y permite el uso de reglas que controlan la introducción de datos no permitiendo aquellos que vulneran alguna de ellas.

En la figura 3.12 puede verse una imagen de SPSS Data Entry. Este programa ofrece un entorno más sofisticado para la construcción de cuestionarios, además de ofrecer una gran integración con el paquete estadístico SPSS. Una opción presente en este sistema es la capacidad de posponer la aplicación de las reglas de análisis hasta haber terminado de introducir los datos, generando un informe indicando los errores encontrados. Esta función puede ser utilizada para analizar la calidad de la introducción de los datos.

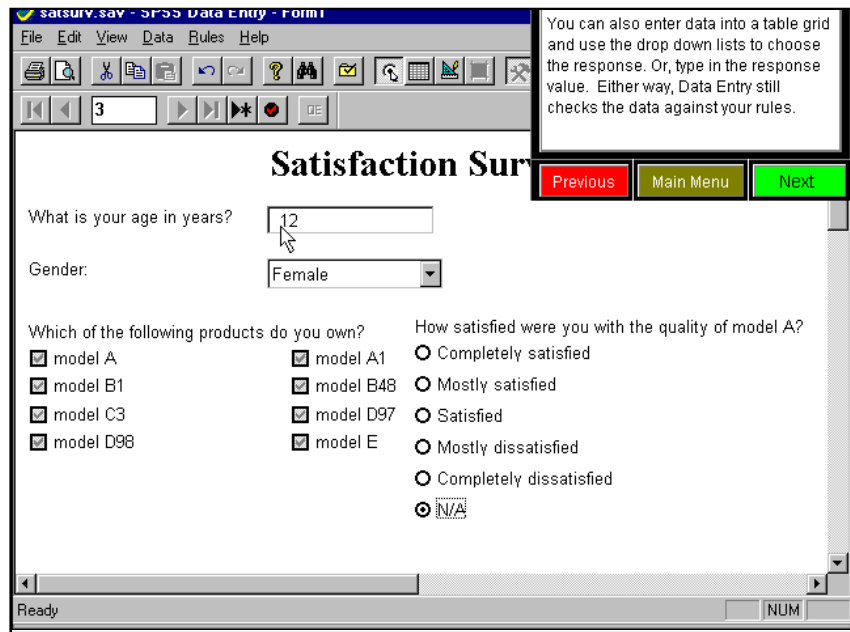
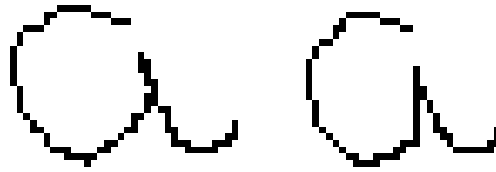


Figura 3.12: El programa SPSS Data Entry

### 3.3. La información es registrada manualmente pero es introducida automáticamente

Los scanners son instrumentos de input que permiten la traducción de información sobre papel al formato digital. Estos aparatos suelen apoyarse en dos elementos:

- Software de reconocimiento de caracteres: Dado que los registros sobre papel presentan mucha variabilidad física (diferentes estilos de escritura, herramientas, etc.) las imágenes digitalizadas de la información necesitan de un cierto grado de interpretación (figura 3.13). Esta ambigüedad ocurre incluso cuando la información ya ha sido producida mediante un método mecánico (por ejemplo un texto en un periódico o una revista) pero se agrava cuando el origen es manual.



*Figura 3.13. Ampliación al doble de tamaño de un intento de escribir la letra a exactamente igual dos veces*

- Papel con marcas ópticas. Este papel está diseñado de tal modo que el software es capaz de determinar los lugares en los que se encuentra los elementos que debe leer el software de reconocimiento de caracteres. Ello facilita enormemente la tarea y aumenta el porcentaje de reconocimiento.

Una pregunta de interés es hasta qué punto el ahorro obtenido mediante este método es lo suficientemente importante como para que realmente valga la pena utilizarlo, en lugar de la introducción manual. Barton et. al. (1991) llevaron a cabo la comparación entre tres métodos de introducción de datos para un estudio con varios cientos de sujetos y unas 10.000 variables. En el primero todo el proceso era manual. En el segundo y en el tercero se utilizaba un scanner, pero mientras que en el segundo se usó un programa que reconocía la información tal y como era tecleada en un cuestionario tradicional, en el tercero se utilizaron hojas especiales. Utilizando estimaciones del tiempo requerido a partir de submuestras del trabajo el esfuerzo necesario aproximado hubiera sido respectivamente: 7514 horas/persona, 2120 horas/persona y 400 horas/persona.

Obviamente, utilizar un scanner hace mucho más rápido el proceso. No obstante, revisando sus resultados es posible ver que, cuando se utilizan métodos automáticos existe una cierta cantidad de esfuerzo inicial que puede no verse compensada cuando se trabaja con conjuntos de datos pequeños.

Las razones para utilizar este tipo de software pueden ser otras además de la económicas. Mattox et. al (1997) justifican la utilización de un sistema de este tipo para hacer más rápida la disponibilidad de datos acerca de incidentes de conducción. De este modo, la utilidad de esta información aumenta ya que los sujetos encargados de su recogida (agentes de tráfico) obtienen feedback más rápidamente y pueden utilizarla en su trabajo. Ello también favorecería en su opinión la motivación acerca de la tarea de recogida de datos por estos sujetos y la calidad de la información así obtenida.

En la figura 3.14 se encuentra una pantalla de Teleform. Un programa especializado en la lectura de datos por medio de scanner.

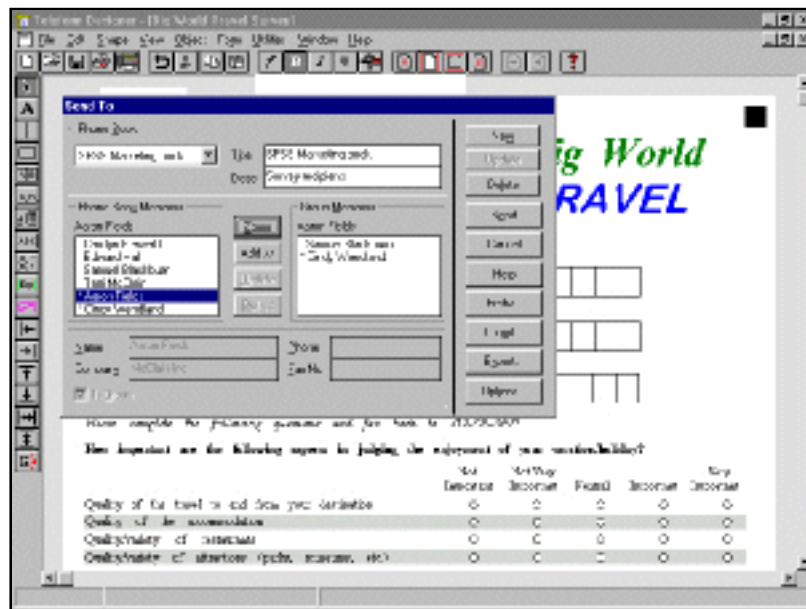


Figura 3.14: Teleform para Windows.

### 3.4. La información es introducida manualmente pero no hay un registro previo

Las herramientas anteriores están dirigidas a solucionar el problema de traducir información en un formato físico a un formato digital. Puede plantearse que una forma de evitar ese problema por completo es introducir directamente la información en el ordenador sin que haya un registro previo en ese formato físico.

Hay varias ventajas de este sistema:

- El salto de preguntas y el llenado automático con datos se puede hacer sin intervención del administrador.
- Se pueden realizar comprobaciones de consistencia y de ese modo mantener informado al entrevistador de los posibles problemas en el mismo momento en que se realiza la entrevista.
- La información puede ser transmitida inmediatamente al centro receptor utilizando un modem.

- La manipulación de los datos puede ser evitada utilizando métodos de encriptación.
- En ocasiones se plantea que pueden existir efectos de orden a la hora de contestar al cuestionario planteándose como solución aleatorizar el orden de presentación de las preguntas. En caso de pruebas de rendimiento la razón para aleatorizar el orden de las preguntas es evitar la copia de las respuestas entre los sujetos realizando la prueba. La aleatorización resulta difícil de hacer con cuestionarios de lápiz y papel y generalmente sólo se hacen un número limitado de versiones para las cuales puede existir igualmente ese efecto de orden. En el caso de la aplicación con ordenador esta aleatorización puede llevarse a cabo con relativa facilidad.

Por otro lado, existen otros inconvenientes:

- En caso que el cuestionario no esté bien diseñado, los entrevistadores pueden tomar menos acciones correctoras que cuando está basado en papel.
- Puede hacer más difícil la toma de datos (el ordenador puede exigir unas acciones más complicadas que el papel o cuestionario).
- Puede limitar los lugares en los que se podrían tomar los datos.

Finalmente, existe tecnología que puede mejorar algunos de los inconvenientes de estos sistemas y que es conveniente conocer:

- Introducción de datos vía teléfonos con generadores de tonos: Es posible utilizar el marcador del teléfono para transmitir símbolos sencillos a través de la línea por lo que se puede disponer una encuesta a través de ellos.
- Reconocimiento de voz: También apropiada para entrevistas telefónicas, resulta necesaria cuando la línea telefónica no admite información digital y sólo es posible transmitir información analógica. Un sistema de reconocimiento de la voz permitirá finalmente determinar las contestaciones realizadas por los usuarios. Este sistema se limita también a códigos sencillos.
- Ordenadores portátiles o Asistentes Personales Digitales. Sobre todo los segundos han sido diseñados de tal modo que la introducción de datos resulta especialmente sencilla. Del tamaño de una agenda, utilizan una pantalla sensible a un lápiz óptico (v. figura 3.15) con el que se pueden rellenar los campos de información o incluso textos por medio de tecnología de reconocimiento de la escritura. Estos aparatos pueden utilizar correo electrónico a través de un teléfono portátil por lo que la

información puede ser transmitida a un centro receptor prácticamente de modo inmediato, acelerando el proceso de recogida de datos de una manera enorme.



*Figura 3.15. Asistente personal con lápiz óptico y teclado suplementario*

Podemos pensar en dos situaciones diferentes que parten de ese planteamiento y que analizaremos por separado (Saris, 1991). En la primera, el entrevistador o persona encargada de recoger los datos tiene a su disposición un aparato electrónico que le permite introducir la información a la vez que la obtiene. En la segunda, es el propio usuario el que tiene a su disposición el sistema electrónico e introduce los datos por sí mismo siguiendo las instrucciones que se le van dando en la pantalla. Veremos en primer lugar la situación en que el cuestionario es administrado por el entrevistador y en segundo la correspondiente al cuestionario autoadministrado.

#### **3.4.1. Cuestionario pasado por el administrador**

Usando la nomenclatura de Saris (1991) existen dos tipos de sistemas que pueden ser encuadrados dentro de esta categoría. Sistemas de Entrevista Telefónica Asistida por Computador (SETAC) y Sistemas de Entrevista Personal Asistida por Computador (SEPAC).

En SETAC el entrevistador se sienta delante de un ordenador y llama a un sujeto para ser entrevistado. Una vez ha empezado el contacto, todas las preguntas van apareciendo en la pantalla y el entrevistador las va leyendo y rellenando a medida que el entrevistador va contestando. Desde el punto de vista del entrevistado no hay diferencia entre una entrevista asistida o no asistida por ordenador. Desde el punto de vista del

entrevistador su tarea se puede ver aliviada al llevar a cabo el ordenador ciertas tareas tal y como control de la consistencia de las respuestas, saltos, etc.

Una ventaja de este sistema es la posibilidad de gestionar la selección de la muestra por el ordenador. A partir de una lista de datos de teléfono de sujetos a entrevistar los programas de ordenador pueden seleccionar aleatoriamente muestras de individuos para entrar en contacto. El programa puede anotar la fecha y la hora en la que se realizó la llamada y el resultado que se obtuvo. En caso de no localizar al sujeto muestreado el programa puede guardar registro del intento y realizarlo más tarde para evitar el sesgo que supondría sólo entrevistar a los sujetos que responden a la primera llamada. También es posible situar límites sobre el número de sujetos de un tipo determinado a muestrear. El entrevistador puede guardar información para cada una de las llamadas en relación con el resultado obtenido o, si esta es completada satisfactoriamente, los datos obtenidos. En caso de estudios longitudinales, en los que el sujeto es contactado en diferentes momentos, el programa puede establecer cuando corresponde hacer la llamada de nuevo. Por último, en caso de tener un grupo de entrevistadores haciendo el trabajo es posible llevar registro de datos acerca de su productividad y rendimiento. También es necesaria una base de datos que centralizaría la información acerca de los teléfonos usados y que distribuiría las nuevas llamadas a los distintos puestos.

El programa denominado MaCATI incluye un módulo denominado *Cleric* encargado de la gestión de los números de teléfono y que está diseñado para funcionar en un ordenador central. Los entrevistadores tienen ante sí algo parecido a la figura 3.16, en la que es posible ver una ventana dedicada a la información acerca de las llamadas telefónicas y otra detrás dedicada a rellenar las contestaciones de los sujetos entrevistados.

La situación es diferente en SEPAC, en ella el entrevistador lleva a cabo las entrevistas cara a cara pero se apoya en un ordenador para registrar las respuestas. En ella, el entrevistador dispone de un dispositivo electrónico portable por medio del cual puede registrar las contestaciones de los entrevistados. Saris (1991) opina que, desde el punto de vista del entrevistado, la situación no sufre grandes variaciones aunque se ha sugerido que la simple presencia del ordenador puede hacer variar la situación de entrevista. Este autor cita otras investigaciones en las que se obtienen las siguientes conclusiones:

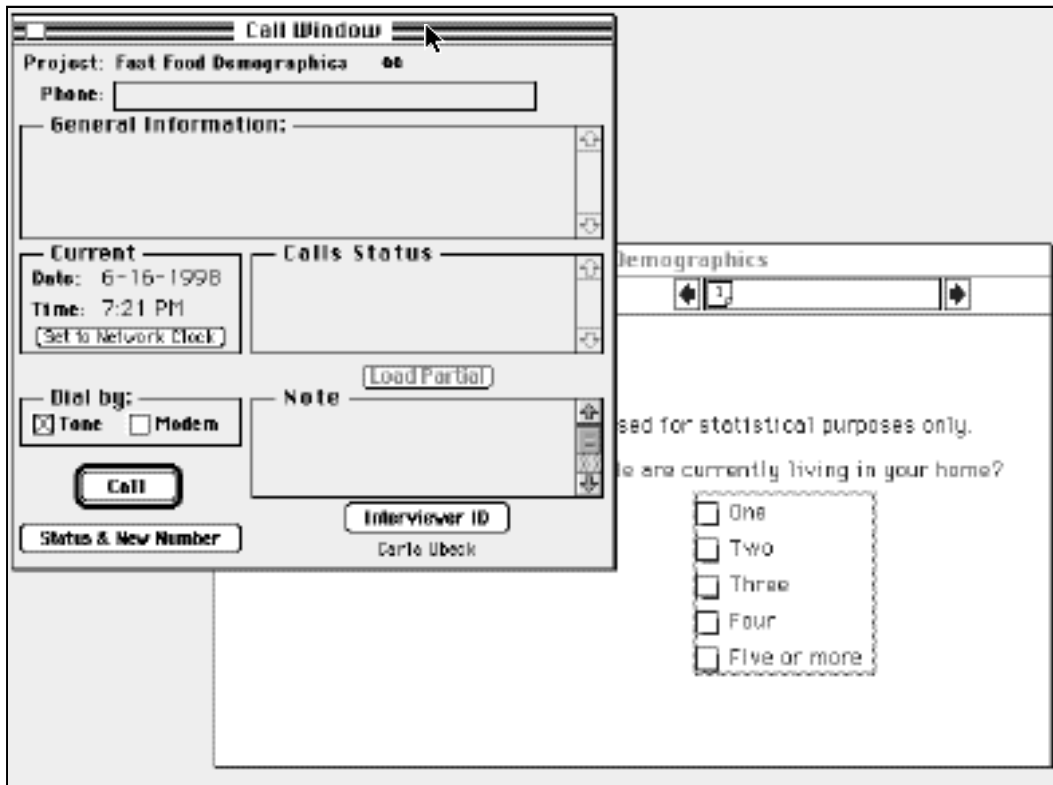


Figura 3.16. MaCati, un programa para la gestión de encuestas por teléfono

- No parecen haber cambios con respecto a los rechazos a contestar o a la no - respuesta incluso con preguntas "delicadas" del tipo de los ingresos.
- Las entrevistas usando ordenador duran más que las realizadas con cuestionarios de lápiz y papel y las notas hechas por el entrevistador son más cortas.

Por último, aunque el entrenamiento en el uso de estos cuestionarios puede no ser muy exigente siempre será mayor que en el caso de los cuestionarios tradicionales. Además, un cuestionario por ordenador puede resultar menos flexible en cuanto a su utilización por lo que determinados casos imprevistos pueden resultar más difíciles de manejar, exigiéndose por tanto un grado mayor de cuidado en cuanto a su diseño.

Una situación intermedia entre esta y aquella en la que los cuestionarios son autoadministrados es cuando los entrevistados contestan las preguntas utilizando el ordenador por sí mismos pero hay un entrevistador disponible para ayudarles en esta tarea. Esta situación presenta la posibilidad de entrevistar a varios sujetos al mismo tiempo por lo que los costes pueden verse reducidos.



Ya sea mediante teléfono o por otro medio las entrevistas que se apoyan en el uso de ordenadores presentan para Saris (1991) una serie de características o problemas comunes:

1) En entrevistas normales, cuando las preguntas o las categorías de respuesta son excesivamente complejas, las preguntas son leídas en voz alta al entrevistado. Cuando el texto es excesivamente complejo los entrevistados pueden optar por pedir que se les repita la pregunta o bien simplemente dar una respuesta cualquiera (generalmente la primera o la última alternativa). Cuando hay muchas categorías de respuesta el sujeto probablemente tendrá problemas para recordar la lista que se le ha leído. En SETAC esto resulta difícil de solucionar mientras que en SEPAC es posible mostrar la pantalla al entrevistado o tener una pila de tarjetas como en las entrevistas sin ordenador. Sin embargo, puede hacerse difícil para el entrevistador manejar el ordenador y la pila de tarjetas simultáneamente.

2) Ayuda por parte del entrevistador. En caso de entrevistas en papel el entrevistador puede ayudar con preguntas o situaciones complicadas. Debido a los patrones de salto y llenado que es posible diseñar en el cuestionario puede ocurrir que cada entrevista parezca completamente diferente al entrevistador. En ese caso, el entrevistador, a menos que se le entrene adecuadamente, puede tener también problemas en entender el cuestionario.

3) Manteniendo el ritmo de la entrevista. Si el tiempo necesario para pasar de una pantalla a otra en el ordenador es excesivo los entrevistados pueden sentirse molestos o aburridos. Muchos aspectos del diseño del cuestionario pueden influir en esa velocidad aunque las prestaciones actuales de los ordenadores hacen muy difícil que una tarea como ésta sea problemática.

4) Determinación de inconsistencias. Puesto que los programas pueden ser diseñados para detectar las inconsistencias en la respuesta a una pregunta, o entre las combinaciones para varias preguntas, el entrevistador puede apoyarse en él para interrumpir la entrevista cuando lo crea conveniente. Lo que el programa de ordenador no puede hacer es obtener la respuesta correcta, por lo que es necesario contrastar esta inconsistencia con el entrevistado. Esto puede producir dos problemas: a) el entrevistador puede encontrar más cómodo rellenar la inconsistencia con un valor cualquiera que reclamarla del entrevistado, y b) los entrevistados pueden sentirse molestos ante esta reclamación.

Por último, señalar que, a pesar de las posibles ventajas que estos sistemas pueden presentar, Saris (1991) remarca el que el esfuerzo necesario para diseñar estos cuestionarios es sin duda mayor. Por ello recomienda una evaluación cuidadosa de las

ventajas e inconvenientes en relación con la situación a tratar antes de tomar la decisión de utilizar este tipo de formularios.

### **3.4.2. Cuestionarios autoadministrados**

En el caso de cuestionarios autoadministrados la situación de entrevista puede decirse que cambia radicalmente (Saris, 1991). El sujeto que responde debe tomar un papel mucho más activo y debe utilizar el ordenador por sí mismo. Si hay un entrevistador presente su tarea se limita a disponer el equipo y a aclarar las dudas que puedan presentarse. Esto presenta la ventaja de que el entrevistador no puede cambiar las preguntas puesto que el entrevistado las lee directamente de la pantalla sin su intervención. Por otro lado, el esfuerzo de diseño necesario para permitir a los entrevistados llevar a cabo esta tarea por sí mismos es mayor ya que el manejo del teclado o del ordenador en general, así como la lectura en pantalla impone dificultades que es conveniente suavizar en la medida de lo posible. Una correcta elección del hardware puede resultar de gran ayuda aquí. Por ejemplo, las pantallas sensibles a la presión (*touchscreens*) son consideradas más intuitivas en la mayoría de los casos mientras que controlar un ratón de ordenador supone serias dificultades a los usuarios sin entrenamiento previo.

Saris (1991) señala las siguientes consideraciones a tener en cuenta en un sistema de este tipo:

1) Preguntas con recuerdo libre. En ocasiones el investigador no quiere dar las alternativas posibles de respuesta al entrevistado sino que prefiere que este conteste aquello que primero le venga a la mente. Este tipo de preguntas suele ser rellenado por el entrevistador en cuestionarios normales utilizando una lista con las respuestas más probables en la cual marca la contestación obtenida. Si la respuesta no está en la lista se puede llevar a cabo una codificación posterior. Las preguntas con recuerdo libre pueden ser utilizadas en cuestionarios autoadministrados pero presentan el inconveniente de que el sujeto tiene que teclear la respuesta, con los consiguientes errores de escritura, con lo que al realizar la comparación con la lista el ordenador puede no identificar la categoría correcta. Para solucionarlo se pueden considerar las aproximaciones a los valores de la lista como válidas. Sin embargo, este método dará resultados inferiores al obtenido cuando un entrevistador se encarga de esta tarea, suponiendo un serio problema si el cuestionario implica saltos en función de esas respuestas

2) Preguntas con inconsistencias. Si el sujeto produce respuestas inconsistentes el programa debería ser capaz de mostrar un listado de los valores potencialmente

incorrectos en pantalla, de tal modo que el entrevistado resolviera el problema a partir de ella. El procedimiento para llevar a cabo esta tarea debería resultar lo más sencillo posible para evitar confundir al entrevistado.

3) Ayudas visuales. En situaciones de entrevista sin uso de ordenador no es fácil proporcionar ayudas visuales. El entrevistador debe cargar una serie de tarjetas que irá mostrando en los momentos adecuados, aumentando de este modo la complejidad de su tarea. Cuando se utiliza un ordenador estas ayudas son mucho más fáciles y, con las capacidades de los ordenadores actuales, es posible utilizar catálogos de gran complejidad. Para preguntas relacionadas con fechas, el ordenador puede proporcionar un calendario en el que, por ejemplo, respuestas anteriores y otros sucesos (fiestas, días de la semana) aparezcan marcadas.

4) Pantallas resumen y de corrección. Aunque los procedimientos de evaluación de las consistencias permiten detectar una cierta cantidad de posibles errores, todavía pueden permanecer aquellos que no violan ninguna de las reglas de consistencia. Si las ramificaciones y los saltos dependen de estas preguntas las consecuencias pueden ser más graves puesto que el sujeto se puede enfrentar a preguntas que le son completamente extrañas o inadecuadas. Para solucionar este problema se pueden proporcionar pantallas de resumen de las contestaciones realizadas por los entrevistados siempre que se vaya a tomar decisiones de navegación por el cuestionario importantes. En ellas los sujetos pueden comprobar sus respuestas y, en caso de que alguna sea incorrecta, dirigirse al lugar en que fue realizada para modificarla.

5) Escalas psicofísicas: En muchas ocasiones resulta interesante que los sujetos contesten sobre algún tipo de escala continua para indicar la fuerza con la que expresan una opinión. En cuestionarios de lápiz y papel la corrección se realiza utilizando mediciones manuales. Esto es laborioso y puede producir errores. Si las contestaciones se llevan a cabo en el ordenador es posible realizar una medición automática mucho más rápida y precisa. Además es posible proporcionar métodos de respuesta más complejos (dibujando líneas de distinta longitud en la pantalla por ejemplo) difíciles de utilizar en un cuestionario de lápiz y papel.

Una de las aplicaciones más interesantes de los cuestionarios autoadministrados es tal y como expone Saris (1991) los estudios longitudinales. En el sistema descrito por él, la recogida de datos se produce mediante una combinación de visitas por medio de entrevistas personales, telefónicas y basadas en un ordenador cedido a los sujetos incluidos en la muestra. Los sujetos son instruidos para responder en el ordenador a una serie de preguntas que son modificadas semanalmente o siguiendo otra periodicidad. Los

datos son transferidos vía modem al centro de recogida de datos. En caso de no responder se puede contactar con ellos telefónicamente o realizar una visita al domicilio para establecer las razones de este hecho. Puesto que los sujetos están acostumbrados a la utilización del equipamiento su manejo no supone los problemas que implica el primer contacto con él.

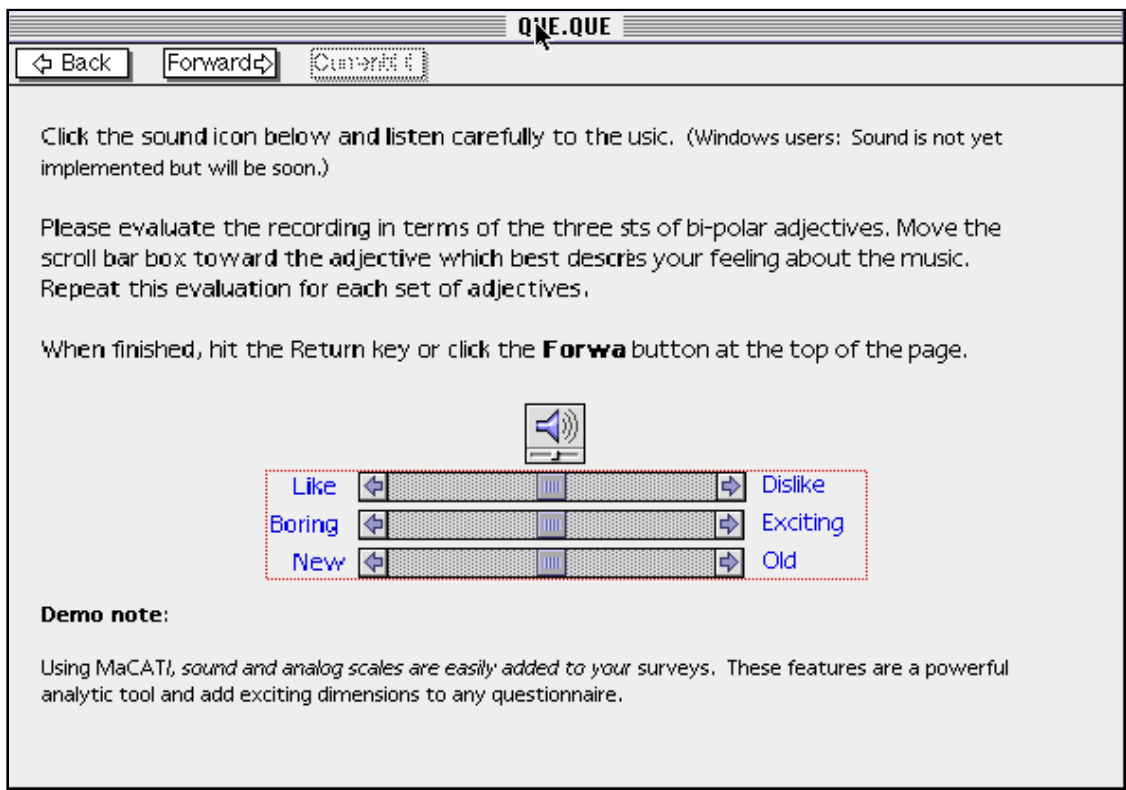


Figura 3.18. Cuestionario autoadministrado elaborado con MaCATI

La situación de cuestionario autoadministrado puede producirse de diferentes maneras:

- El entrevistador dispone de un lugar especial con uno o varios ordenadores al que los entrevistados acuden para realizar la entrevista.
- El programa se encuentra en un disco de ordenador que es suministrado al usuario y que tiene que utilizar por su cuenta.
- El programa se encuentra en un lugar público tal y como unos grandes almacenes en un pequeño kiosco.

- Por medio de la red Internet se puede pedir a los usuarios que visiten un determinado servidor que rellenen un cuestionario. Esta información puede combinarse con el registro automático del comportamiento del sujeto (qué partes del servidor visita, cuanto tiempo dedica a cada parte, etc.).

El programa de ordenador MaCATI permite la creación de cuestionarios para autoentrevistas que pueden ser contestados por medio de un ordenador normal o utilizando la Internet. Este programa admite muchos tipos diferentes de preguntas y respuestas tal y como sonidos o gráficos. Posibilita hacer comprobaciones de consistencia y realiza saltos siguiendo las contestaciones de los entrevistados. Un programa separado permite mostrar los cuestionarios en pantalla a los entrevistados. En la figura 3.18 se muestra una pantalla en la que el entrevistado puede oír el sonido de una canción haciendo click con el ratón en el icono del altavoz. Su opinión acerca de tres variables es recogida utilizando las barras de desplazamiento de la parte inferior.

#### **3.2.4. Consideraciones de diseño de pantallas**

La introducción de datos en un ordenador es una tarea que, a pesar de su sencillez, incorpora dificultades para los usuarios que pueden hacer disminuir la calidad de los datos. Pantallas mal diseñadas, procedimientos de introducción de datos incómodos y poco prácticos, excesiva monotonía, etc. son sólo algunos de los problemas que pueden producir esa falta de calidad. Métodos aparentemente novedosos (uso de voz, pantallas táctiles, etc.) pueden asimismo parecer como ideas interesantes en un principio pero, luego, la experiencia práctica demuestra que las ventajas supuestas a menudo no llegan a concretarse.

Una fuente de información fundamental para el diseñador de procedimientos de introducción de datos es el libro "Guidelines for designing user interface software" por S. Smith y J. Mosier (1986). Este libro describe reglas que un diseñador debería tener en cuenta a la hora de elaborar software. Este libro es gratuito y puede examinarse gratuitamente en la Internet en la dirección <<http://www.sydney.csiro.au/hci/guidelines/sam/guidelines.html>>. Estas reglas han sido extraídas a partir de la literatura sobre el uso de ordenadores por lo que el diseñador interesado puede consultar las fuentes y las razones tras cada recomendación. En la tabla 3.19 se indican las secciones que incluye este libro y el número de reglas para cada una de ellas.

Secciones	NºReglas
• Introducción de datos	199
• Presentación en pantalla	298
• Secuencia de introducción	184
• Guía del usuario	110
• Transmisión de datos	83
• Protección de datos	70

Tabla 3.19: Secciones y número de reglas en el Smith y Mosier (1986)

Tal y como los mismos autores advierten, el uso de conjuntos de reglas de este tipo no garantiza obtener un producto perfecto, aunque su uso debería mejorar su calidad o, al menos, su consistencia con otros productos similares. Existe una versión electrónica y es posible realizar búsquedas acerca de temas o palabras clave que se relacionen con el problema que se esté manejando en cada momento. Como un ejemplo se incluye la regla 1.0/5 incluida en la primera sección (tabla 3.20):

## 1.0/5 Un método único para introducir datos

Diseñar la introducción de datos y las pantallas asociadas para que el usuario pueda seguir con un método de entrada de datos y no tenga que cambiar a otro muy a menudo.

**Ejemplo:** Minimizar los cambios entre el *lightpen* (un instrumento con el que se marca en la pantalla para introducir datos) y el teclado para introducir datos.

**Ejemplo negativo:** Un usuario no debería cambiar de un teclado a otro, o de un ordenador a otro para realizar partes de una tarea de introducción de datos.

**Comentario:** Esta regla asume que la tarea de introducción de datos se lleva a cabo de modo intenso y con mucha sobrecarga, por lo que la eficiencia en la introducción de datos es necesaria.

**Referencia:** BB 2.11, EG 6.1.1, Foley Wallace 1974, Shneiderman 1982 (estas abreviaturas se explican en el propio libro).

**Ver también:** 1.1/14 (Esto señala a a otra regla dentro del libro).

Tabla 3.20: Un ejemplo de regla en el Smith y Mosier (1986)

En la tabla 3.20 se puede ver en primer lugar la descripción de la regla, y luego dos ejemplos de su aplicación, un comentario acerca de su ámbito de aplicación, las referencias de artículos o libros relevantes para justificar esta regla y finalmente otras reglas relacionadas.





# ***Transferencia de archivos de datos***

## **4.1. Introducción**

Generalmente sucede que los programas de gestión y análisis de datos son muy similares, y, tras haber optado por uno de ellos, lo utilizaremos para realizar la mayor parte de las tareas. Sin embargo, puesto que ningún programa es capaz de cubrir todas las tareas (hay quien dice que los mejores son los que no lo intentan), de vez en cuando, necesitaremos utilizar características específicas de otro. Por otro lado, otras personas con las que colaboramos pueden haber optado por programas diferentes al nuestro, así que, en caso que queramos hacer uso de sus datos o resultados, tendremos que pasarlos a traducirlos al formato de nuestro programa. Una última situación que necesita de importación/exportación es cuando utilizamos como fuente datos archivados por instituciones, hospitales, empresas, u otras organizaciones que utilicen sistemas de gestión propios cuya conexión con los programas estadísticos no está resuelta. En este caso, las transferencias de archivos pueden ser una tarea bastante compleja que puede requerir una gran cantidad de esfuerzo y conocimientos.

Hablaremos de importación para referirnos al proceso que nos va a permitir abrir/leer un fichero de datos que no se encuentre guardado en el formato del programa que estemos utilizando. Hablaremos de exportar datos cuando los datos que estamos utilizando en un programa son convertidos a un formato diferente al que éste está acostumbrado a utilizar. Importación y exportación son ambas operaciones de transferencia de archivos.

Realizar traducciones es necesario debido a que cada programa utiliza un *formato* diferente. Puesto que cada situación puede requerir una cierta disposición de los datos, cada programa opta por la que encuentra más apropiada. Una hoja de cálculo, por ejemplo, necesita ser capaz de detectar rápidamente las celdas interconectadas entre sí, para, de ese modo, realizar los cambios oportunos cuando una de ellas cambia. Un paquete estadístico en cambio puede utilizar una serie de resúmenes estadísticos de la tabla de datos que le permite realizar nuevos cálculos con mayor rapidez.

Realizar transferencias entre archivos con formatos diferentes es una gran fuente de problemas. A menudo, las interpretaciones no se realizan correctamente y, por ejemplo, una variable toma el lugar de otra, los valores faltantes adquieren otro significado, o los decimales dan lugar a números erróneos. Afortunadamente, existen gran cantidad de programas que facilitan estas tareas, además de algunas convenciones relativamente sencillas, que permiten solucionar las situaciones más comunes.

Revisaremos en primer lugar algunos de los formatos más comunes y luego examinaremos los métodos y programas para hacer traducciones entre ellos. Esta sección asume que los datos traducidos tienen una forma de tabla, ya que, aunque existe la posibilidad de traducir información que tiene estructuras más complejas, en este último caso lo más aconsejable es remitirse a la propia documentación de los sistemas implicados.

## **4.2. Formatos**

Podemos distinguir dos tipos de formatos. Los primeros son más simples y más apropiados para el intercambio de datos. Los segundos son más complejos, más dependientes del programa o sistema que los originó y, aunque en ocasiones es posible utilizarlos para el intercambio de datos, resultan menos universales.

#### 4.2.1. Codificación de caracteres

El formato ASCII (*American Standard Code for Interchange of Information*-Código Americano Normalizado para el Intercambio de Información) es el sistema básico para la representación de caracteres (Sanmartín, et al., 1990). Este código deriva de los teletipos, los cuales necesitaban un método sencillo para enviar códigos alfabéticos y otros especiales (cambios de línea, de párrafo, etc.). La versión original utilizaba 128 mensajes diferentes (2<sup>7</sup>) y otra ampliada pasó a tener 256 (2<sup>8</sup>), aunque los nuevos 128 códigos están menos normalizados y presentan problemas de interpretación (sobre todo en lo que respecta a caracteres propios de determinados lenguajes tal y como acentos, etc.).

Hay que advertir que este formato es normalmente una versión simplificada de la información y que pueden perderse muchos de los atributos que ésta tuviera. Por ejemplo, los cambios de estilo del texto (negrita, cursiva, etc.) no se conservarán en el formato ASCII. Sólo los caracteres, los cambios de párrafo y, en ocasiones de línea, permanecerán.

Naturalmente, los números también pueden ser representados como caracteres y en la mayoría de los casos este será el método preferido. Los números presentan la ventaja de existir un alto grado de acuerdo en cuanto a su representación mediante el código ASCII por lo que no ocurren problemas de interpretación (otra cosa es cuando los números van acompañados de otros añadidos como veremos más adelante).

Para crear un archivo en formato ASCII podemos utilizar multitud de programas que son capaces de guardar la información con la opción sólo texto o texto ASCII o similar. Los *editores de texto* son programas sencillos diseñados con este propósito, pero tanto los *procesadores de texto* como *hojas de cálculo* o *bases de datos* suelen ser capaces de crear este formato.

#### 4.2.2. Codificación de tablas de datos en formato ASCII

Puesto que los datos normalmente vienen en estructuras rectangulares no basta con transferir los valores sino que es necesario una serie de convenciones que permitan reconstruir la organización en filas y en columnas de la tabla original.

Existen fundamentalmente dos formatos:

a) Formato libre o delimitado: En el formato libre cada valor está separado del siguiente por un carácter especial (generalmente, un espacio en blanco, una coma o, más usualmente en los últimos tiempos, un símbolo de tabulador). Generalmente el final de

cada párrafo está también marcado por un símbolo de cambio de párrafo. En la figura 4.1 es posible ver un ejemplo de datos en formato libre. En él se han marcado con este símbolo ¶ las separaciones entre los valores. El cambio de párrafo se ha marcado con este otro ¶. En la parte de arriba se incluye la cabecera con los nombres de las variables, la cual sigue las mismas convenciones que los propios datos. Esta cabecera no es un componente propio de esta tabla por definición, pero muchos programas la admiten y resulta cómoda

SEXO	GENERO	HISTORIA	MATEMATICAS
1	1	2	3
2	2	3	4
3	1	1	1
4	1	1	1
5	2	1	2
6	2	1	1
7	2	1	1
8	1	1	1
9	2	1	2
10	2	1	1

Figura 4.1. Datos en formato libre

Una variante de este formato es el libre con una sola columna o fila. SPSS por ejemplo admite este tipo de formato para una columna. Los datos de la figura 4.1 se dispondrían como en la 4.2.

1	1	2	3	2	2	2	3	2	3	1
2	1	1	4	1	2	1	1	3	2	1
3	1	6	2	4	4	7	2	1	1	1
4	1	1	2	1	4	2	3	1	2	2
5	2	2	1	1	1	1	1	1	1	1

Figura 4.2. Datos en formato libre con una sola fila de datos

Los datos ocupan varias filas por el efecto de texto en el que son mostrados pero el símbolo de cambio de párrafo ¶ no aparece hasta el final por lo que los datos estarían dispuestos como si ocuparan una fila. Para leer estos datos correctamente es necesario indicar el número de variables al programa que los va a utilizar. De este modo hará la separación en líneas correctamente.

De entre los dos formatos se recomienda el primero ya que es más similar en apariencia a la tabla que se quiere construir. Esto facilita detectar errores.

b) Formato fijo: 'En el formato fijo los datos correspondientes a cada caso no aparecen separados por caracteres específicos (tabulaciones, espacios en blanco, comas, etc.), sino que el valor correspondiente a cada variable aparece en una posición predeterminada, la misma para todos los casos. Dada esta condición, las variables no necesitan un carácter que las separe entre sí. Por otro lado, en las instrucciones para leer la información es necesario indicar exactamente cuándo empieza y cuándo termina una columna para evitar errores. Además, hay que indicar la parte que corresponde a decimales y la parte entera en caso de tener datos de este tipo. Los datos anteriores aparecerían así (aquí los hemos separado en filas para facilitar su lectura aunque estrictamente no es necesario). Por ejemplo, en estos datos la variable SUJ ocupa las dos primeras columnas, la variable GENERO ocupa la cuarta y la variable HISTORIA las columnas 6 7 y 8. Como es posible ver, algunas columnas han quedado vacías. Esto se ha hecho para mejorar la legibilidad, pero no es estrictamente necesario. También, los puntos decimales han sido dispuestos para coincidir en las mismas columnas, rellenando con ceros los datos sin decimales. Para apreciar la organización de estos datos resulta imprescindible utilizar un tipo de letra monoespaciado (en el que todos los caracteres tienen el mismo tamaño). En el ejemplo se ha utilizado Courier.



1:1	2:5	2:0	2:0	2:0	2:0	2:0	2:0	2:0	2:0
2:2	2:5	2:0	2:5	2:0	2:5	2:0	2:5	2:0	2:5
3:1	2:9	1:9	1:9	1:9	1:9	1:9	1:9	1:9	1:9
4:1	2:5	1:0	1:0	1:0	1:0	1:0	1:0	1:0	1:0
5:2	3:0	1:0	3:5	1:0	3:5	1:0	3:5	1:0	3:5
6:2	2:5	4:5	4:0	4:0	4:0	4:0	4:0	4:0	4:0
7:2	1:8	1:0	1:8	1:0	1:8	1:0	1:8	1:0	1:8
8:1	2:5	1:5	4:0	1:5	4:0	1:5	4:0	1:5	4:0
9:5	2:5	1:0	2:5	1:0	2:5	1:0	2:5	1:0	2:5
10:2	2:0	1:0	1:0	1:0	1:0	1:0	1:0	1:0	1:0

Figura 4.3. Formato fijo

#### 4.2.3. Codificación de números

El formato ASCII no es especialmente interesante desde un punto de vista de optimizar los recursos del ordenador. El formato *binario* en cambio presenta la ventaja de poder ser buscado y analizado más rápidamente si se utiliza el software apropiado. Sin embargo, no es posible interpretarlo directamente por medio de caracteres y suele ser muy

dependiente del programa o máquina que lo generó. Mientras que el número 5 necesita 7 bits para ser almacenado en formato ASCII, esto puede reducirse en formato binario a sólo 3 (aunque normalmente se utilizan 4). Muchos paquetes estadísticos utilizan internamente formatos binarios (ICPSR, 1997).

#### **4.2.4. Otros formatos más específicos**

Aunque el resto de los formatos que vamos a discutir tienen de algún modo la característica de estar ligados a ciertos programas concretos, tienen (o han tenido) el suficiente protagonismo como para merecer ser discutidos aunque sea brevemente.

- **Formatos portables:** Tanto SPSS como SAS incorporan versiones de sus formatos binarios que permiten el paso de sus datos entre diferentes ordenadores. Esto permite usar los mismos datos con el mismo paquete estadístico (SPSS o SAS) en diferentes ordenadores.
- **Formato BASIC:** BASIC es un lenguaje de programación famoso por permitir obtener resultados rápidos aunque no muy elegantes. El formato de datos de BASIC está basado en caracteres. En él, los campos están separados por comas, los de tipo texto están entrecomillados y cada párrafo está separado del otro por retornos de carro.
- **Formato SYLK:** Es un formato basado en caracteres. Está especialmente dirigido a hojas de cálculo (es decir a tablas). Lo más esencial es que señala a cada celda mediante sus coordenadas x e y. Este formato es más apropiado para hojas de cálculo ya que en ellas puede haber muchas celdas sin valores. Puesto que muchos usuarios utilizan estos programas para introducir datos puede ser un formato interesante para ellos.
- **Formatos de otros programas:** A continuación sigue una lista de algunos de los formatos que muchos programas son capaces de exportar e importar: SPSS, SAS, ACCESS (una base de datos de Microsoft), DBASE (un programa de base de datos que fue en su momento el más común), EXCEL (quizás la hoja de cálculo más popular en este momento), LOTUS (la que fue la hoja de cálculo más popular hasta recientemente), etc. Conocer estos formatos puede resultar de gran utilidad a sus usuarios, aunque la lista es muy larga y puede cambiar según las modas o momentos en que se elabora. Así, es aconsejable consultar las capacidades de importación/exportación de los programas correspondientes en el momento en que se emprendan estas tareas.

#### 4.2.5. Programas de traducción de datos

Existen al menos dos programas especializados en realizar traducciones de datos entre diferentes programas de manejo de datos. DBMS/Copy y Stat-transfer. El primero de ellos es distribuido por SPSS y realiza traducciones entre 80 formatos de programas de análisis, gestión y representación de datos diferentes. Stat-Transfer por otro lado afirma traducir tantos formatos como DBMS/Copy y ser más rápido. En la figura 4.4 se muestra una pantalla de este último.

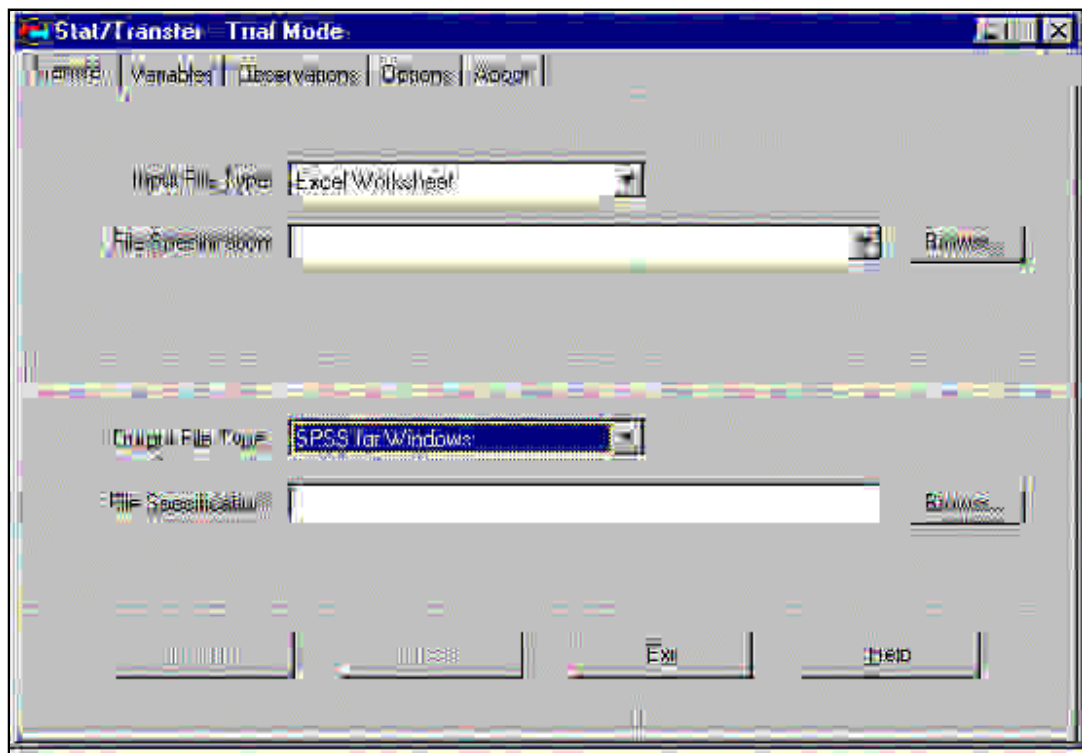


Figura 4.4. Una pantalla de Stat-Transfer





# 5

# ***Control de la Calidad de Datos***

## **5.1. Introducción**

Dentro de las posibles comprobaciones de la calidad de los datos distinguiremos entre los métodos dirigidos a comprobar la fidelidad de los datos y los que se refieren a examinar su consistencia o coherencia.

Los primeros permitirían determinar si el dato introducido es correcto con respecto a la fuente de la que provienen sin analizar si éste es aceptable o no. Los segundos se dirigirían a detectar valores que, aunque posiblemente iguales con respecto a la fuente original, son inconsistentes desde un punto de vista lógico.

Por ejemplo, en un cuestionario en el que se pregunta sobre el género y la posición en la familia un sujeto tiene como contestaciones "hombre" y "esposa". Una comprobación del primer tipo examinaría si el valor introducido en la base de datos es correcto respecto al cuestionario rellenado originalmente. Si esta comprobación da como resultado que la respuesta referida a la relación familiar fue introducida correctamente, y

que la verdadera contestación a la primera pregunta es "mujer", el valor en la base de datos sería modificado de una manera directa.

Sin embargo, si el cuestionario original tiene marcadas realmente las contestaciones "hombre" y "esposa", nos encontraríamos ante un error de consistencia. El problema estaría ahora en determinar cual de las contestaciones es incorrecta ("hombre" podría ser "mujer" y "esposa" podría ser "marido"). Este tipo de inconsistencia es denominada lógica o determinística. Otro tipo de inconsistencia no determinística es cuando los valores son lógicamente posibles pero abiertamente inusuales, hasta tal punto que despertarían incredulidad. Por ejemplo, un sujeto que ve la televisión 24 horas al día estaría en esta categoría u otro que declarara como profesión "obrero no especializado" y unos ingresos altos despertarían un sentimiento de incredulidad que nos llevaría con toda seguridad a comprobar sus afirmaciones para certificar su veracidad. Este tipo de datos han sido denominados en la literatura estadística "outliers" y su detección puede ser utilizada en el contexto del diagnóstico de la calidad de los datos tal y como será mostrado en una sección posterior.

Veremos a continuación métodos para realizar estos dos tipos de comprobaciones.

## **5.2. Control de la fidelidad**

Los siguientes métodos pueden utilizarse con este propósito:

a) Método de la ejecución duplicada: Los datos ya introducidos son revisados por dos operadores diferentes con respecto a un criterio, produciendo un registro de 0 y 1 para cada ítem. Cero indicaría datos incorrectos y uno correcto. Los desacuerdos entre el operador  $A_1$  y el operador  $A_2$  se resolverían reexaminando los datos originales (West and Winkler, 1991). Mediante este método se garantiza que aquellos ítems en los que se coincida en el valor de 0 y los que haya desacuerdo serán reexaminados o tratados adecuadamente. West y Winkler (1991) aplican métodos bayesianos para la estimación de errores restantes en la base de datos no detectados (es decir, que los dos operadores produjeron un registro de 1) tanto para el caso del supuesto de tasas de error iguales o diferentes.

b) Método de los errores conocidos: En él, la base de datos es ampliada para incluir ítems adicionales que se sabe que son erróneos. La base de datos es comprobada entonces por un único individuo para saber cuántos errores detecta. Esto da información

acerca de la ejecución del individuo, la cual se puede extrapolar a los errores desconocidos. West y Winkler utilizan métodos bayesianos igualmente para estimar la cantidad de errores que permanecerán no detectados a partir de esa información.

c) Método de la introducción doble: En él la información es introducida por dos operadores diferentes y un programa de ordenador es el encargado de producir un registro de 0 y 1 indicando si el valor es igual o diferente. En realidad, es similar al método de la ejecución duplicada pero en aquel la tarea de los operadores era comprobar cada valor con una fuente que indicaba si el valor era correcto o no, mientras que en éste, los sujetos introducen la información y es el ordenador el que determina su concordancia. Programas diseñados específicamente para la introducción de datos tal y como el SPSS Data Entry o BMDP Data Entry tienen opciones para hacer esas comprobaciones (una hoja de cálculo puede también realizar esta tarea con suma facilidad). De nuevo, existe la posibilidad que los dos sujetos cometan la misma equivocación en el mismo ítem y el error permanezca no detectado.

En nuestra opinión, el último de los métodos tiene mayor aplicabilidad en situaciones de recogida de datos mediante cuestionarios y posterior introducción manual. Los otros dos parecen más adecuados para situaciones de control de calidad en las que dos evaluadores tienen que decidir sobre la calidad o aceptabilidad de los ítems. En caso de aplicarse a un cuestionario, el efecto de fatiga en la tarea de revisión es seguramente mucho mayor que en la de introducción de datos. Por ello, puede ser interesante hacer que los sujetos reintroduzcan la información ya que la propia tarea les mantendrá activados. En cualquier caso, cada situación concreta puede dictar la utilización de un método u otro.

### **5.3. Control de la consistencia**

La consistencia de los registros de una base de datos puede ser comprobada mediante dos tipos diferentes de pruebas (Naus, 1982, Naus, et al., 1972).

- a) Exactas (también llamadas por definición, determinísticas o lógicas); y
- b) Aproximadas (empíricas o probabilísticas).

En las primeras, las comprobaciones se dirigen a eliminar aquellas inconsistencias que violan algún tipo de regla implícita o explícita para los datos que estamos manejando. Por ejemplo, un varón no puede padecer ciertas enfermedades, cierto tipo de profesiones tienen salarios entre ciertos límites, etc.

Las segundas se dirigen a examinar aquellos puntos que, aún no rompiendo ninguna regla lógica, comprobamos que se comportan de modo diferente al del resto de los puntos. Por ejemplo, un sujeto contesta en un cuestionario sobre valoración de ciertos productos de modo completamente inverso al resto. Aunque sus contestaciones pueden ser legítimas desde un punto de vista lógico podemos sin embargo dudar de ellas y contemplar la posibilidad de una mala interpretación del sentido de las preguntas.

La primera parte de esta sección estará dedicada a las pruebas denominadas exactas y la segunda a las empíricas.

### **5.3.1. Pruebas exactas o determinísticas**

En la sección dedicada a la introducción de datos hemos señalado que los programas más adecuados para esta tarea suelen incluir facilidades para evaluar los valores y decidir si son admisibles. Esta tarea puede realizarse en dos momentos diferentes: En el momento en que se está realizando la propia introducción o posteriormente, cuando todos los valores han sido introducidos y se desea hacer una evaluación de su calidad.

Los procedimientos aquí explicados pueden ser utilizados independientemente del momento en que se apliquen. Obviamente, si no va a resultar posible volver a las fuentes originales una vez se ha detectado una inconsistencia, lo mejor es realizar la comprobación en el mismo momento en que se recoge la información. No obstante, si las contestaciones son sobre un cuestionario en papel esto no es fácil de hacer.

Para Fellegi y Holt (1976) la *edición de datos*, término con el que designan la tarea de examinar la corrección de unos datos implica fundamentalmente las dos tareas siguientes:

- a) Comprobar cada campo en cada registro de una encuesta (la respuesta registrada a cada pregunta en el cuestionario) para diagnosticar si contiene una entrada válida; y
- b) La comprobación de entradas en ciertas combinaciones predeterminadas de campos para diagnosticar si las entradas son consistentes unas con otras (Fellegi and Holt, 1976).

Entre las comprobaciones del tipo a tendríamos las siguientes :

- 1) Cuando el conjunto de valores admisibles es finito se comprobaría si el valor que se intenta introducir pertenece a él. Por ejemplo, género del individuo sólo admitiría los valores 0 y 1. Valores de 2 ò 1003 serían considerados inconsistentes.

2) Cuando el conjunto de valores admisibles es infinito y numérico pero existe un límite esencial o simplemente razonable, se comprobaría si el valor cae dentro de esos límites. Por ejemplo, altura de un individuo menor que 2.30 y mayor que 1.30 (si hablamos de sujetos adultos).

3) Cuando los valores admisibles tienen que ser obligatoriamente diferentes entre sí (por ejemplo cuando se trate de una variable identificadora) se puede comprobar que el valor introducido no está ya presente en algún registro de la base de datos.

En general, las comprobaciones del tipo a se derivarían directamente de la estructura y los códigos del cuestionario mientras que las de tipo b necesitarían de un conocimiento más amplio de la materia tratada en la encuesta. Este segundo tipo de comprobaciones constituyen la parte principal de esta sección. Algunos ejemplos de comprobaciones de tipo b son las siguientes:

a) Cuando el conjunto de valores posibles para unas variables es finito resulta posible hacer comparaciones del tipo SI...ENTONCES. Por ejemplo, si el sexo de un sujeto es varón su número de embarazos es cero.

b) Si el conjunto de valores posibles para unas variables es infinito podemos en ocasiones encontrar límites condicionales entre sí. Por ejemplo, si el número de empleados en una empresa es un valor dado, la suma de horas trabajadas diariamente para todos los empleados en una empresa no puede superar un límite dado.

c) En ocasiones, ciertos valores pueden encontrarse que son la suma (o cualquier otra operación aritmética) de otros, por lo que, aunque podría utilizarse el ordenador para realizar el cálculo y acortar de este modo la recogida de datos, podemos pedir esa información y comprobar que el resultado coincide. Por ejemplo, la nota media del curso tiene que ser igual a la suma de las asignaturas dividido por el número de éstas. En caso de haber diferencias, algún valor debe ser incorrecto.

La utilización de estas técnicas de comprobación permite detectar inconsistencias en la información. Cuando una de éstas aparece Fellegi y Holt (1976) listan las opciones que, en principio, tenemos disponibles:

1) Comprobar los cuestionarios originales con la esperanza que la información allí estará correctamente especificada y los fallos provengan de la fase de transcripción al ordenador.

2) Contactar con el sujeto que produjo la respuesta y preguntarle de nuevo.

3) Utilizar a personal entrenado para eliminar las inconsistencias usando ciertas reglas.

4) Usar un ordenador para lo mismo que en 3.

5) Eliminar los registros con inconsistencias.

Obviamente, en caso de ser posible las dos primeras opciones son las mejores (aunque costosas). En otro caso las opciones 3 a 5 tendrán que ser utilizadas. De entre ellas, la 3 y la 4 son muy similares, aunque Fellegi y Holt (1976) se inclinan por esta última ya que los ordenadores tienen muchas ventajas respecto a los operadores humanos: menos fatiga, más consistencia en la aplicación de las reglas y, seguramente, menor coste.

La opción número cinco es un tipo de borrado de información semejante al método denominado *listwise* y que será expuesto con más amplitud en el apartado dedicado al análisis de valores faltantes. En pocas palabras, esta opción implica una pérdida de información con, muy posiblemente, efectos negativos.

Una vez detectada una inconsistencia en un registro y seleccionada la opción cuatro el problema consistiría en lo siguiente:

a) Determinar qué campo concreto o campos es el responsable de hacer que la comprobación lógica haya fallado. Volviendo a un ejemplo anterior, si un sujeto tiene como valor en el campo género "hombre" y como valor en el número de embarazos "2", existen dos posibles causas para la inconsistencia. Bien el género correcto es "mujer" o bien el número de embarazos es "0". Sin una fuente externa a la que consultar la solución será en muchos casos indeterminada y será necesario determinar criterios que seleccionen la pregunta más apropiada, tal vez siguiendo criterios lógicos (Fellegi and Holt, 1976) o de tipo probabilístico (Naus, et al., 1972).

b) Una vez determinado el campo que se desea corregir, determinar qué valor puede ser más aceptable. Este es un problema de asignación, el cual presenta dos situaciones diferentes. Cuando tratemos con variables de tipo cuantitativo es posible utilizar los métodos descritos en la sección dedicada a valores faltantes de este texto, con la complejidad adicional de garantizar que los valores estimados satisfagan las limitaciones lógicas correspondientes. Cuando tratemos con variables de tipo cualitativo, la situación es un tanto diferente a la tratada en el apartado de valores faltantes y será descrita en esta sección.

Existen dos métodos aplicables en este caso aunque ambos pueden ser puestos en relación.

a) El método de Fellegi y Holt. Este método considera que el campo responsable de que un registro no sea consistente es aquel que aparece en todas las pruebas lógicas, explícitas o implícitas, que ese registro no haya podido pasar. El siguiente ejemplo (Naus, 1982) permitirá explicar esto.

Supongamos que tenemos las siguientes tres variables:

V1) Fecha en la que un miembro de la familia utilizó por primera vez cierto servicio social.

V2) Fecha en que el cabeza de familia utilizó por primera vez el servicio social.

V3) Fecha en que un miembro de la familia utilizó por última vez el servicio.

Estas tres variables tienen dos comprobaciones que pueden derivarse directamente (y que llamaremos explícitas).

$$\text{I) } V1 \leq V2$$

$$\text{II) } V2 \leq V3$$

A partir de estas reglas podemos derivar una regla implícita que sería:

$$\text{III) } V1 \leq V3$$

La última regla es considerada como implícita porque puede ser derivada de las otras dos y porque no puede incumplirse independientemente de las otras dos.

Supongamos que tenemos el siguiente caso: V1: 1979, V2: 1980, V3: 1978. Este caso no cumple la regla II ni la regla III. Si examinamos esas reglas veremos que V3 aparece en ambas por lo que resulta razonable pensar que el valor correspondiente a esa variable en este caso es el incorrecto.

En este caso cualquier valor igual o superior a 1980 sería aceptable y podría ser elegido para V3. Sin embargo, esto es quizás excesivamente ambiguo por lo que sería necesario tener en cuenta otros criterios si no queremos limitarnos a elegir un valor al azar de entre los aceptables. El método probabilístico precisamente hace referencia a la utilización de información externa a la que está produciendo a la que está produciendo la inconsistencia, tanto para su identificación como para su corrección.

b) El método probabilístico de Naus y otros (1972). El método anterior puede generar muchas situaciones en que el número de variables implicadas en todas las pruebas lógicas falladas sea mayor que uno. En ese caso la decisión sobre la variable a corregir puede apoyarse en métodos probabilísticos que funcionen del siguiente modo. Examinando los registros de la base de datos que han pasado las pruebas lógicas podemos establecer la frecuencia condicional, y de aquí la probabilidad condicional, de asociación entre los valores sospechosos y no sospechosos en el caso que estemos considerando. Aquellos valores sospechosos de estar incorrectos pero que en el resto de la base de datos presentan una alta asociación con los valores no sospechosos recibirían una probabilidad de ser correctos mayor que aquellos que, además de ser sospechosos, tuvieran una baja asociación con la parte del registro que no parece estar equivocada.

Un ejemplo de esta aproximación sería el siguiente. Tenemos tres variables:

V1: Profesión (una lista de aproximadamente 40 categorías).

V2: Género (hombre o mujer).

V3: Número de embarazos (0, 1, 2, ó +3).

Podemos derivar fácilmente la regla  $V2=\text{hombre} \rightarrow V3=0$ .

Tenemos un caso con los siguientes valores: V1: administrativo, V2: hombre, V3: 2. Este caso no pasa la comprobación por lo que asumimos que uno de los campos es incorrecto. Supongamos que para el resto de los casos que sí pasan la comprobación se comprueba que la profesión administrativo tiene una proporción de mujeres muy alta (digamos el 90%). En este caso podríamos utilizar esta información para desconfiar del valor en V2 mucho más que del valor en V3.

Este método también nos proporciona una manera de decidir por un posible valor a utilizar para corregir el error. Partiendo de la información disponible podemos realizar una estimación de qué valor sería más probable en esa situación y consecuentemente asignarlo. En ese caso, necesitaríamos tener en cuenta las precauciones con respecto a la asignación que se detallan de modo más amplio en la sección dedicada a datos faltantes. Esta estimación podría hacerse utilizando por ejemplo los modelos para datos categóricos descritos en Schaffer (1997) o quizás utilizando redes neuronales (Gharamani and Jordan, 1994).



### 5.3.2. Valores inusuales pero no errores determinísticos

Una vez introducidos los datos en el ordenador y obtenido un archivo que puede ser enviado a un programa de análisis es conveniente realizar una inspección previa de ellos (Afifi y Clark, 1984; Tabachnik y Fidell, 1989). Esa inspección en general cubre varios aspectos tal y como comprobar supuestos estadísticos (normalidad, homoscedasticidad, etc.) así como otros relacionados con el proceso de datos (veracidad de los datos, valores faltantes, etc.). Esta inspección ha recibido diversos nombres tal y como *data editing* (edición de datos), *screening* (monitorización), *laundering* (limpiado), validación o control de la precisión del input (Naus, 1982). En este apartado nos centraremos en las comprobaciones relacionadas con la fidelidad de los datos con el objetivo de disminuir al máximo los posibles *errores de datos* existentes en ellos. Antes de ello introduciremos algunos conceptos que nos pueden ayudar a comprender el resto del material aquí tratado.

Una primera consideración es la referida al origen de los errores. Estos pueden haberse producido en diversos lugares:

a) En la fase de recogida de datos: Hojas de registro mal organizadas, preguntas ambiguas, engaño deliberado, etc. pueden producir que los datos tengan errores.

b) En la introducción de los datos: Siempre y cuando los datos sean introducidos manualmente es posible que se produzcan errores humanos. No obstante, los avances en este campo (programas de lectura de datos) pueden en un futuro cercano reducir este componente de error enormemente.

c) En las transformaciones y manipulaciones de los datos: Muchas de las transformaciones de los datos implican un componente de programación, el cual es susceptible, naturalmente de error.

Barnet y Lewis (1994) distinguen entre tres fuentes de variabilidad que pueden ser encontrados en nuestros datos:

a) Variabilidad inherente a los datos: Esta es la expresión de la forma en que las observaciones varían de forma natural en la población. (Valores extraños que surgen como consecuencia de una gran variabilidad de este tipo son denominados *outliers*).

b) Error de medida: Cuando se toman medidas del objeto de estudio a menudo el instrumento utilizado impone una variabilidad como si fuera un factor inherente. También el redondeo al tomar medidas o las equivocaciones de registro. Este error puede ser reducido usando ciertos métodos.

c) Error de ejecución: Esta fuente de error proviene de haber llevado a cabo una recogida de datos imperfecta (Valores extraños que surgen como producto de un gran error de tipo b) o c) son denominados valores espúreos o simplemente errores).

Una clasificación de las formas en que podemos intentar detectar los errores que nos resulta interesante para esta exposición es la que diferenciaría entre *directos* y *estadísticos*.

a) Métodos directos: Serían todos aquellos que se basarían en una comprobación de la corrección de los datos de modo individual. Por ejemplo, alguien podría revisar los datos uno por uno comprobándolos con los registros originales, o podría muestrear los registros que comprueba. Otro método podría ser el de la doble introducción de datos.

Los métodos directos pueden ser aplicados no solamente a los datos sino también a ciertas propiedades de los datos. Por ejemplo, los máximos y/o los mínimos de ciertas variables pueden ser conocidos, así como los valores admisibles cuando se trate de variables categoriales. También es posible elaborar una serie de pruebas lógicas que determinen si existe coherencia entre los diversos valores asignados a un caso (Naus, 1982). En general, los métodos directos aplicados a las propiedades de los datos pueden ser incorporados al propio proceso de introducción de los datos, de tal modo que, el error aparezca en el mismo momento que se intenta introducir. Por ello, estos métodos han sido descritos previamente y no serán comentados aquí de nuevo.

b) Llamamos métodos estadísticos a aquellos destinados a mostrar inconsistencias o valores extraños pero no lógicamente imposibles. Estos métodos están destinados fundamentalmente a detectar *outliers* (valores fronterizos o desplazados) antes que errores. Sin embargo, puesto que algunos *outliers* serán en realidad simples errores de datos estas técnicas son aplicables aquí. La distinción entre errores de datos y *outliers* es lo suficientemente importante como para justificar una explicación más detallada que será emprendida en la sección posterior.

Los métodos estadísticos también permiten detectar errores que afectan a toda la variable en su conjunto. Por ejemplo, aunque la temperatura es una variable que no se encuentra limitada en cuanto a sus valores superiores y puede superar el valor de 30 o 40 grados Celsius en ocasiones, sería difícil creer que la temperatura *media* para un lugar en, por ejemplo, España fuera de 40 grados. Así, si situáramos un control sobre esta variable en función de su máximo razonable es posible que individualmente ninguno de sus valores lo superara pero en su conjunto se produjera un resultado poco creíble.

### La distinción entre outliers y errores de datos

Para Barnett y Lewis (1994) un *outlier* es una observación o subconjunto de observaciones que parece ser inconsistente con el resto de los datos. Más tarde describiremos con más detalle esta definición pero antes enfatizaremos sus diferencias con el concepto de simple error de datos.

Si al examinar unos datos llegáramos a la conclusión que uno de ellos parece un *outlier* podríamos, como primer paso, comprobar los datos originales y decidir si estos coinciden con los valores que tenemos actualmente registrados. En caso que no sea así y supuesto que los datos originales suponen un grado de veracidad mayor que los actualmente registrados podríamos simplemente cambiar el valor incorrecto y proceder con más comprobaciones. Barnett y Lewis (1994) proporcionan algunos ejemplos en los que cambios en la escala de medida, errores de edición de datos, etc. dan lugar a este tipo de situaciones. En caso que *este outlier no pudiera considerarse un error* nos encontraríamos ante un problema de índole más cercana al análisis estadístico que al del proceso de datos previo a éste, entrando en los bien conocidos temas de estimación robusta (Hoaglin, et al., 1983) y/o de identificación de *outliers* (Barnett and Lewis, 1994) que no serán tratados aquí. La figura 5.1 muestra de modo gráfico esta situación. A través de ella podemos ver que los métodos basados en la detección de *outliers* sólo pueden aspirar a detectar una parte de los errores. Precisamente aquellos que, al ser cometidos, los convierten en inconsistentes.

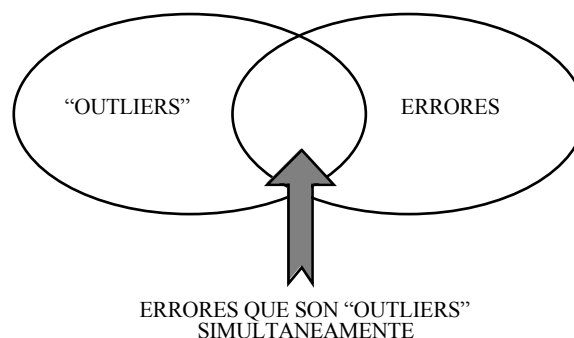


Figura 5.1. Solapamiento entre errores y outliers.

Barnett y Lewis (1997) discuten los modelos estadísticos que explican la producción de outliers. Si tenemos unas observaciones que provienen de una distribución F podemos tener unas cuantas mezcladas en nuestros datos que provienen de una distribución G con unos parámetros diferentes de F. En este caso hablaremos de las

puntuaciones provenientes de G como *contaminantes*. Algunas de las hipótesis utilizadas en la literatura acerca del origen de las puntuaciones G son:

a) Origen determinístico: En este caso las puntuaciones contaminantes son producto de errores de medida o de registro graves. En estos casos, el *outlier* debería ser eliminado o sustituido por el valor correcto.

b) Variabilidad inherente: Es posible que en la variabilidad que nosotros hemos atribuido a F no sea correcta y que en realidad esta sea mayor de la que anticipamos inicialmente. En este caso lo apropiado sería realizar pruebas comprobando si nuestro supuesto inicial acerca de la distribución de la población es correcto o es necesario cambiar de supuesto.

c) Mezcla de poblaciones: En este caso nos podríamos encontrar con que algunas puntuaciones provienen de una población diferente a la representada en el modelo básico.

d) Desplazamiento de puntuaciones: Al parecer el modelo más común para contaminación. En él se establece que todas las puntuaciones provienen de la misma población F salvo unas pocas que aparecen de una versión modificada de F con respecto a sus parámetros. Mucho del trabajo publicado se refiere a la distribución normal.

*Métodos univariados: Métodos directos de estimación del error de datos.*

Partiendo de la situación en que uno dispone de un archivo de datos con posiblemente ciertos errores, el método más directo posible para encontrar y corregir éstos podría ser simplemente examinar los datos uno por uno después de haber producido un listado razonablemente organizado (sábana de datos). Actualmente muchos programas y paquetes estadísticos son capaces de proporcionar esos listados razonables incluso en la propia pantalla, un lujo que no es excesivamente antiguo por otro lado. Listar todos los datos es una opción posible cuando su número no es muy grande o cuando los errores buscados son de un calibre muy grande (por ejemplo, para comprobar el resultado de transformaciones sucesivas de los datos). No obstante, fuera de este contexto éste resulta un camino excesivamente impracticable y poco rentable. Wilkinson (1989) menciona una técnica relativamente simple que permite mejorar el simple listado de los datos y proporcionar cierta capacidad (limitada) de examinar nuestros datos en busca de errores. Se trata de ordenar en función de una variable o varias variables nuestros datos. En el ejemplo de la tabla 5.2 vemos que un error que en los datos de la parte izquierda es difícil de encontrar resulta más sencillo de ver en la parte derecha tras haberlos ordenado (el último caso está repetido).

<i>X1</i>	<i>X2</i>	<i>X3</i>	<i>X4</i>	<i>X5</i>	<i>X1</i>	<i>X2</i>	<i>X3</i>	<i>X4</i>	<i>X5</i>
25	160	172	93	49	22	115	131	54	58
22	138	197	152	43	22	115	181	59	60
22	115	181	59	60	22	138	197	152	43
22	190	190	117	41	22	150	233	176	42
22	115	131	54	58	22	190	190	117	41
25	160	172	93	49	23	154	194	79	49
22	150	233	176	42	24	185	155	89	45
23	154	194	79	49	25	160	172	93	49
24	185	155	89	45	25	160	172	93	49

Tabla 5.2. Al ordenar los datos es posible detectar algunos errores

De todos modos, este camino es, no obstante, limitado puesto que es necesario encontrar la/s variable/s que producirán una ordenación que permita detectar errores de importancia (en el ejemplo de Wilkinson esto resultó fácil, el objetivo era detectar un fraude en unas elecciones y bastó con ordenar por nombres y apellidos para encontrar los repetidos).

Una sugerencia relativamente sencilla útil para estimar el error de datos es seleccionar al azar un grupo de datos originales y comprobarlos individualmente. Según el número de errores podremos llegar a estimar el número de errores totales y obtener una idea de su calidad.

Otro método consiste en introducir dos veces los mismos datos, de tal modo que, en caso de producirse diferencias, se pueda establecer cuándo éstas son producto de errores de introducción. BMDP (Dixon, 1992) incluye un programa que monitoriza esta segunda introducción de datos para determinar si existen diferencias con la primera y genera un archivo dando indicaciones acerca de los lugares en los que se encontraron diferencias. Este método no obstante sólo previene errores de introducción de datos y no aborda la cuestión de aquellas ocasiones en que el dato pueda haber sido incorrectamente recogido.

#### *Métodos univariados: Métodos estadísticos.*

Como hemos anticipado anteriormente, estos métodos están fundamentalmente dirigidos a detectar *outliers* y por tanto su objetivo es más amplio que simplemente el de

captar los errores de datos. No obstante, muchos de los *outliers* detectados serán atribuibles finalmente a simples errores en los datos por lo que existe un cierto solapamiento entre ambos mundos sobre el que nos centraremos en estas páginas. Por otro lado, cuando nos encontremos en la situación en la que existen pocas constricciones lógicas en nuestros datos (límites no establecidos, relaciones no restringidas) los métodos aquí tratados serán la mejor opción disponible en la mayoría de los casos.

Como veíamos anteriormente Barnett y Lewis (1994) encuentran que el aspecto fundamental en la definición de outlier es que parece ser inconsistente con el resto de los datos. Esta definición vemos que se apoya en la idea de que es posible discernir una *coherencia* en los datos disponibles que algún dato está rompiendo y por ello puede ser catalogado como *outlier*, lo cual depende, en definitiva, de la definición de coherencia que adoptemos para cada situación. Esta definición de coherencia es en pocas palabras un modelo que aplicamos a nuestros datos y respecto del que el *outlier* se desvía con claridad. Una modificación de nuestro modelo respecto de los datos puede llevarnos a considerar que un *outlier* no lo es en absoluto. En la siguiente figura se muestra un ejemplo de este extremo. El histograma representa unos valores simulados de la distribución gamma con media igual a 1 (una distribución que suele utilizarse en ocasiones para describir variables tal y como el sueldo o los beneficios). Superpuesta se encuentra una distribución normal con los mismos parámetros que los datos simulados. Si nuestra idea original es que los datos siguen una distribución normal nos encontraríamos ante valores que podríamos considerar *outliers* en el lado derecho del gráfico. También, nos encontraríamos con una característica sospechosa tal y como es la falta de valores negativos. Todas estas consideraciones desaparecerían cuando decidiéramos cambiar nuestro modelo de los datos por uno más cercano a la forma que presentan los datos.

Obviamente, existen valores que son errores y *no aparecen como outliers*. Centrarnos en estos presenta el interés de que esos valores son los que potencialmente están afectados por los errores más grandes y, por tanto, son los que distorsionarían en mayor medida los análisis. Aquellos errores *que no aparecen como valores extremos* podemos pensar que la diferencia con su valor correcto es menor en tamaño y, esperamos, en efecto. También podría ocurrir que un *outlier* inherente a los datos haya sido transformado en otro mucho más aceptable como producto de un error de ejecución de medida. Esta situación puede ser bastante usual, con entrevistadores u otros intermediarios corrigiendo los valores inusuales a otros mucho más aceptables para evitar los problemas provenientes de éstos.

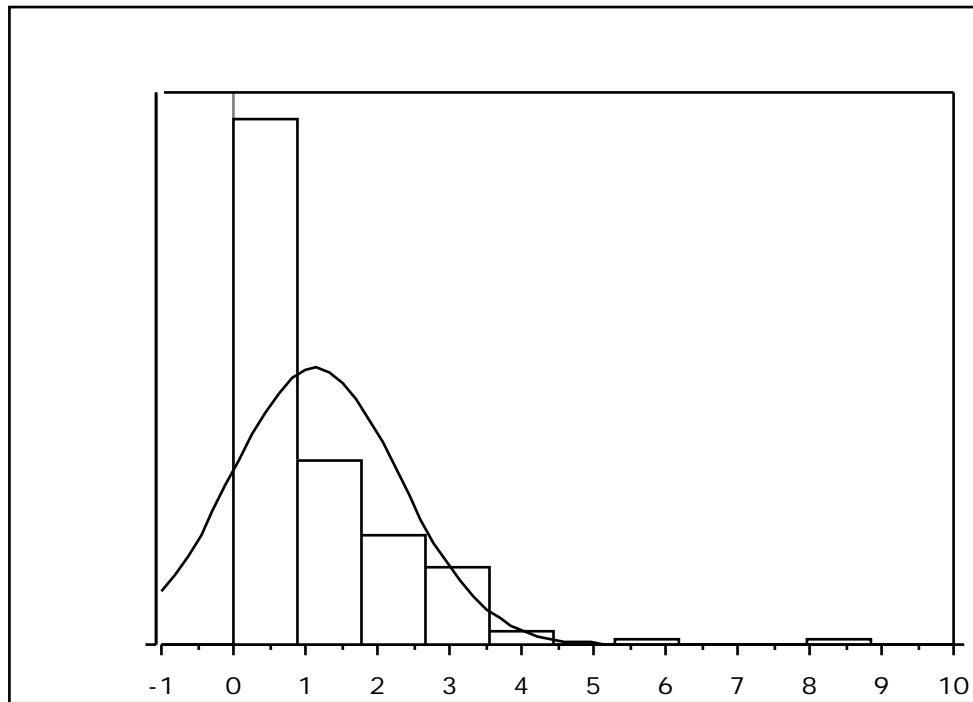


Figura 5.3. Comparación de una muestra de datos siguiendo la distribución gamma (en el histograma) con la normal (superpuesta)

Un concepto interesante es el de distribuciones que puede estar más o menos afectadas por *outliers* (Green, 1982).

#### *Estadísticos descriptivos para Outliers univariados*

Si los datos que están siendo examinados poseen propiedades bien conocidas, los estadísticos descriptivos pueden proporcionarnos información acerca de posibles errores. En la tabla 5.4 es posible ver algunos estadísticos que la mayoría de los paquetes estadísticos suelen ofrecer. Comentaremos a continuación su interés desde el punto de vista de la depuración de datos.

- **Media:** Una media extraordinariamente alta o baja puede ser indicativo de una equivocación de amplitud de los datos. Es decir, un valor extraordinariamente alto o bajo ha sido introducido inadvertidamente. No obstante, el interés de la media para este propósito es muy limitado ya que sólo valores proporcionalmente muy grandes o pequeños la afectarán lo suficiente como para resultar destacable. Por contra, un error sistemático en los datos (con valores excesivamente altos o bajos) pueden ser detectados mediante la media a pesar que los métodos dirigidos a detectar problemas en valor concreto no son capaces de hacer sonar la alarma.

• Número de valores numéricos, no-numéricos y número total de casos: El número total de casos es una información que debería ser contrastada siempre, ya que es un indicador de equivocaciones muy usuales, tal y como omitir a un sujeto/caso o realizar una traducción equivocada entre distintos formatos. No obstante, también es necesario tener en cuenta si los valores introducidos son del tipo esperado, numérico o no numérico. Si en una variable que debería ser exclusivamente numérica aparece algún valor no numérico ello es indicativo de algún tipo de error. En el ejemplo de la tabla 5.4 vemos que la variable Idioma sólo tiene 115 casos frente a los 117 que tienen las otras variables. Dado que estos datos corresponden a las notas en varias asignaturas de los mismos sujetos sería necesario examinar qué es lo que ocurre con esos dos casos de diferencia. Por otro lado, en la variable Lenguaje aparece un valor no numérico, lo cual puede ser también una equivocación.

• Desviación típica: La desviación típica podría indicarnos si los valores de la variable considerada exceden habitualmente de los márgenes esperables. No obstante, sufre los mismos inconvenientes que la media.

• Mínimo y máximo de los datos. La información acerca de los valores extremos es seguramente la más crítica en relación con la depuración de datos. Como indican Barnett y Lewis (1994) un outlier univariado es siempre un extremo. En nuestro caso, la variable Lengua tiene un máximo de 7 mientras que el resto de las variables lo tiene en el 4 o en el 5 (diferencias que a su vez sería interesante comprobar). Por otro lado, la asignatura Historia tiene un valor excesivamente bajo y poco frecuente a la hora de evaluar rendimiento.

Var.	Medias	Numérico	NoNumer	Caso	Desv. T.	Mínimo	Máximo
VALENC	2	117	0	117	1	1	4
LENGUA	2	116	1	117	1	0	7
IDIOMA	2	115	0	115	1	0	5
HISTORIA	1	117	0	117	1	-25	4
MATEM	2	117	0	117	1	1	5
C.N.	2	117	0	117	1	1	5

Tabla 5.4. Estadísticos descriptivos indicando errores de datos

•Valores faltantes ("missing").: Aunque los valores faltantes serán tratados en una sección posterior resulta de todos modos interesante en esta fase comprobar hasta qué punto los valores faltantes en nuestros datos son o no legítimos.



*Métodos gráficos para outliers univariados*

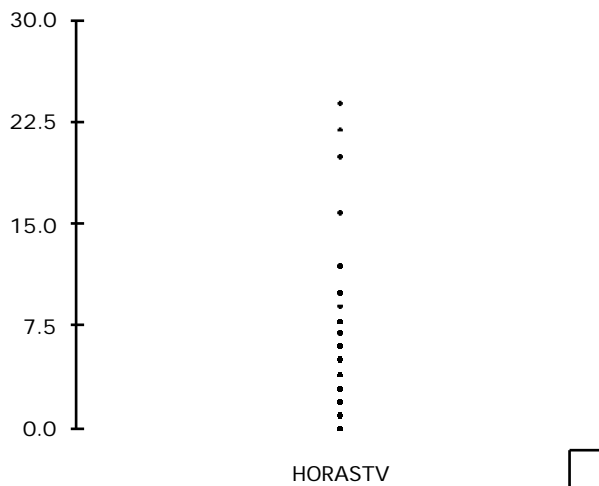
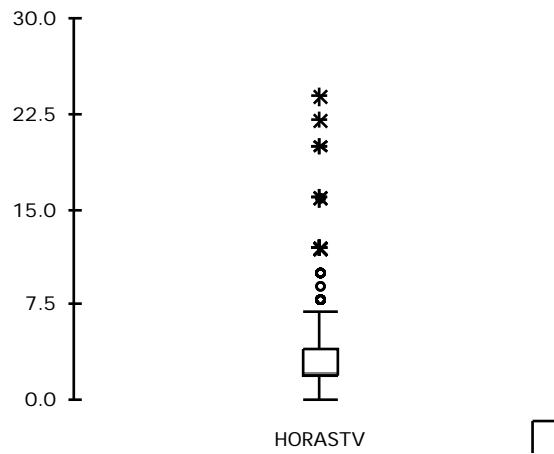
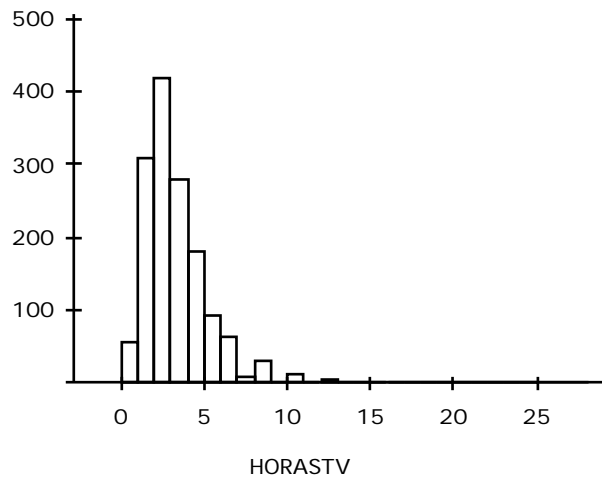
La detección de outliers nos lleva al análisis de valores individuales, tarea para la que los estadísticos descriptivos de la tabla 5.4 no parecen apropiados. En línea con las tendencias de análisis de datos más actuales (Barnett y Lewis, 1994; Cleveland and McGill, 1988; Hoaglin, et al., 1983) los gráficos parecen más apropiados para esta tarea. De este modo podemos captar si un dato es extraño con respecto al resto de los valores. En el siguiente ejemplo se muestran datos de 1500 sujetos con respecto al número de horas diarias viendo TV<sup>1</sup>, representados por medio de tres gráficos bien conocidos: Un histograma, un diagrama de cajas y bigotes y un diagrama de puntos. En general los tres cumplen su propósito de mostrar que parece haber al menos un sujeto que emplea !24 horas diarias en ver la TV!

Ahora bien, aunque 24 horas es tan extremo que posiblemente sólo puede atribuirse a un error no existe una imposibilidad lógica en él. Por otro lado, aunque este valor es el más llamativo y roza lo imposible existen otros sospechosos. Por ejemplo, parece haber algunos sujetos con valores en la misma variable de alrededor 22 horas. Esta situación corresponde al fenómeno descrito por Barnett y Lewis (1994) de *enmascaramiento*. En esta situación un *outlier* muy extremo oculta a otros valores que también lo podrían ser *si ese outlier no existiera*. Este problema afecta fundamentalmente a pruebas de discordancia pero tiene su equivalente en los gráficos.

Una forma de examinar con rapidez los *outliers* y luchar contra el fenómeno del enmascaramiento es utilizar gráficos dinámicos (Cleveland and McGill, 1988, Tierney, 1989, Velleman, 1995, Young, 1996) para examinar nuestros datos. Estos permiten una visualización rápida de la información asociada y asimismo reconstruir el gráfico atendiendo a la información de interés. Data Desk (Velleman, 1995) permite por ejemplo definir una variable ficticia activa que controla los puntos que son seleccionados en el gráfico. Se puede elegir que sólo se muestren los puntos seleccionados, de tal modo que

---

<sup>1</sup> Estos datos son un ejemplo proporcionado por SPSS v. 7.5.



*Figura 5.5. Tres gráficos representando la distribución univariada de la variable horas diarias viendo la TV para una muestra de 1500 sujetos*

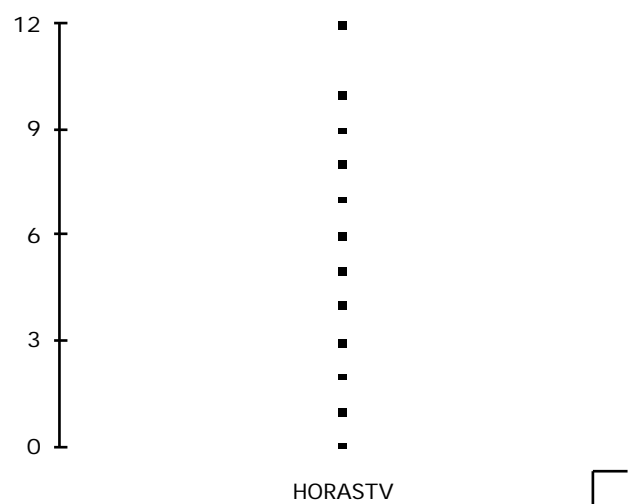
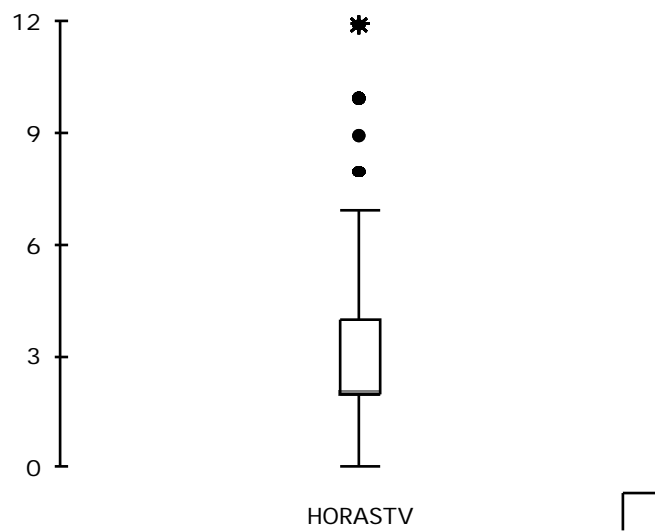
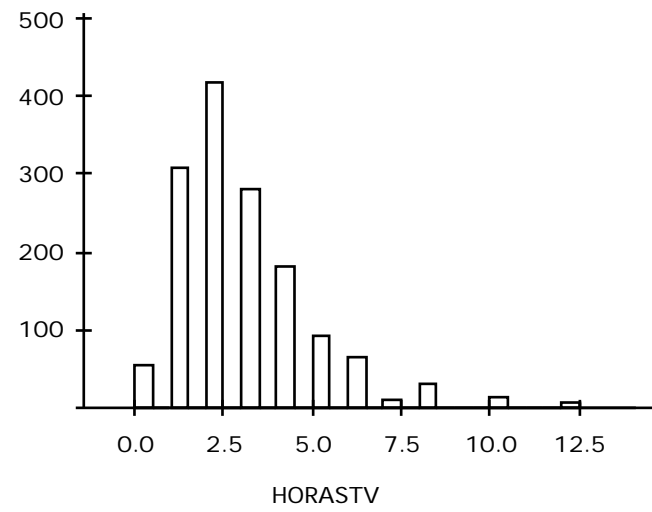


Figura 5.6: Tres gráficos representando la distribución univariada de la variable horas diarias viendo la TV para una muestra de 1500 sujetos una vez eliminados los valores extremos más llamativos

en una acción del ratón el gráfico se reconstruye automáticamente, permitiéndonos centrarnos en los puntos interesantes. Los gráficos en la figura 5.6 fueron conseguidos por medio de un sólo movimiento de ratón y nos permiten examinar la parte central de los datos. Como es posible ver todavía aparecen valores que, enmascarados por los otros valores extremos no se revelaban como sorprendentes, pero que al enfocar nuestra atención sobre ellos aparecen también como dignos de atención.

#### *Pruebas de discordancia para outliers univariados*

¿Cuándo podemos parar con el proceso de examinar los *outliers*? Como hemos visto este proceso puede resultar bastante largo puesto que es habitual que desechar uno o varios *outliers* resulta muy a menudo en el descubrimiento de otros nuevos bajo la luz del nuevo contexto que aparece. Esto nos lleva a la paradoja de un proceso en el que sucesivamente recortamos los valores de los lados de la distribución hasta llegar a un conjunto de valores que parecen adecuados desde el punto de vista del modelo que hemos asumido implícita o explícitamente pero que puede tener poco que ver con nuestro punto de partida. Existe aquí una tensión entre nuestro interés en eliminar valores ilegítimos por un lado, y la fidelidad a la información obtenida realmente por otro. En nuestro caso, esta tensión es menor puesto que nuestro interés principal está en reparar errores y la conducta que corresponde cuando aparece un *outlier* es examinar la información original en la medida de lo posible y corregir los fallos que se hayan presentado. No obstante, si hablamos en términos económicos y de esfuerzo, resulta necesario también tener una regla que nos permita decidir cuándo cesar de seguir examinando *outliers* y aceptar los valores restantes como válidos.

Barnett y Lewis (1994) discuten una gran variedad de tests de discordancia para *outliers* ocurriendo en muestras provenientes de diferentes distribuciones (normal, gamma, lognormal, exponenciales, uniforme, Poisson, binomial, etc.) bajo diferentes supuestos (media conocida o desconocida, desviación típica) y diseñados para realizar evaluaciones individuales o en bloque (es decir, probar si varias puntuaciones pueden considerarse *outliers* simultáneamente). Estas pruebas pueden ser utilizadas para tomar decisiones acerca de qué puntuaciones merecen ser examinadas detalladamente en caso de duda. Un ejemplo para una situación bastante común es la prueba de un *outlier* único (no un bloque de ellos o varios consecutivamente) en una muestra normal con  $\mu$  y  $\sigma$  desconocidas.

$$T = \frac{x_{(n)} - \bar{x}}{s}$$

En donde  $x_{(n)}$  es el valor a probar. Una fórmula equivalente es:

$$T' = \frac{x_{(n)} - \bar{x}'}{s'}$$

En esta segunda fórmula  $\bar{x}'$  y  $s'$  son calculadas no utilizando el valor en prueba  $x_{(n)}$ . Ambas son fórmulas equivalentes. Barnett y Lewis (1994) dan valores críticos al 5% y al 1% y advierten que la costumbre de considerar valores indicativos de *outlier* de 3 o 4 es equivocada ya que puede demostrarse que el máximo de esta ecuación depende del tamaño muestral  $n$  según la siguiente relación:

$$\max T = \frac{x_{(n)} - \bar{x}}{s} = \frac{(n-1)}{\sqrt{n}}$$

#### *Métodos para Outliers multivariados*

La idea de *outlier* puede ser llevada a la situación multivariada. Esto conllevará ciertas particularidades que examinaremos a continuación.

La definición de Barnett y Lewis (1994) según la cual un outlier es un valor que destaca del resto de una manera sorprendente sigue siendo válida aplicada a este caso. No obstante, esa capacidad de sorpresa ya no depende exclusivamente del valor en una variable, sino que puede ser producto de la combinación o relación entre variables. Por ejemplo, puede que un peso de 120 kilos y una altura de 1'40 m no sean valores que, por separado, denominaríamos *outliers* en un conjunto de datos acerca de un grupo de sujetos. Sin embargo, si ambos valores se dan combinados ese valor puede considerarse como un *outlier* (figura 5.7).

De este ejemplo se puede derivar una característica que diferencia el caso multivariado del univariado con respecto a los *outliers*. Mientras que en el caso univariado un *outlier* estaba siempre en los extremos de la distribución (ya fuera en la parte inferior o en la superior) y por tanto bastaba con ordenar los datos para detectar aquellos sospechosos, en el caso multivariante no hay un criterio "natural" que nos permite ordenar los datos en función de su cercanía con una frontera. En la práctica, no obstante, es posible elaborar criterios de centralidad multivariante (por ejemplo, el centroide de los datos calculado como el punto medio simultáneo) a partir del cual definir distancias que nos den idea de la lejanía con respecto a ese centro. Por otro lado, la

complejidad inherente a la visualización gráfica de datos multivariados (Chambers, et al., 1983; Klinke, 1997; Young, 1996) supone una limitación a su inspección.

Un caso que ha despertado mucho interés en los últimos años ha sido el del diagnóstico de *outliers* en situaciones con datos estructurados del tipo del modelo general lineal tal y como en análisis de regresión (Velleman, 1995) y o análisis de varianza, (Hoaglin, et al., 1991). Estos temas no serán tratados aquí de modo exhaustivo ya que parece más apropiado que sean discutidos dentro del contexto de diagnóstico de los análisis realizados mediante estas técnicas.

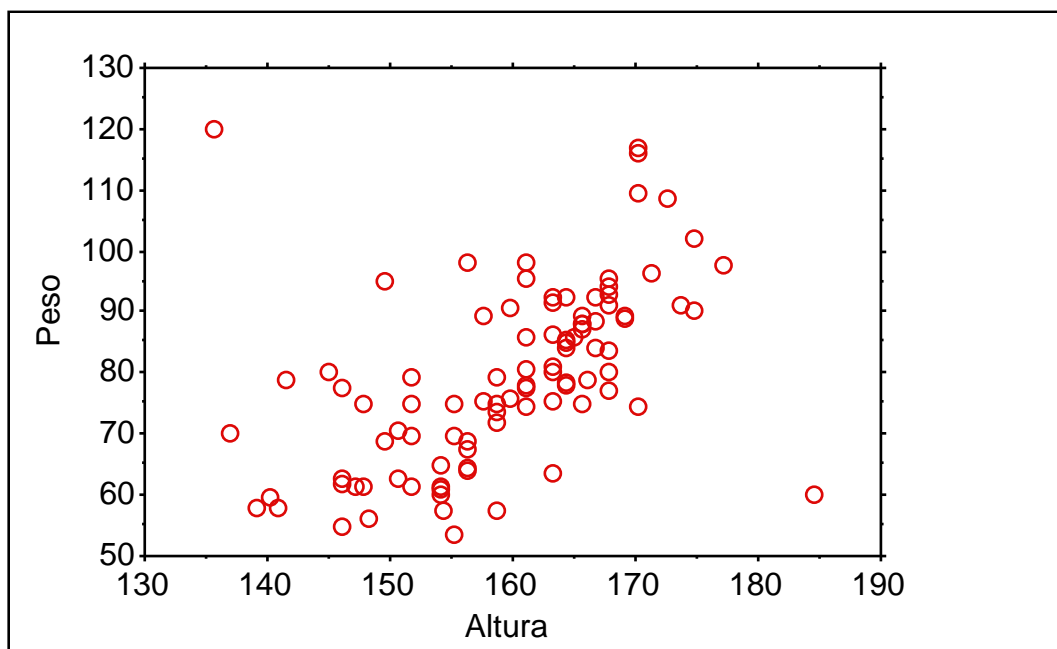


Figura 5.7: Diagrama de dispersión con outliers

#### *Métodos para outliers multivariados*

Revisaremos los siguientes tipos de métodos: Gráficos, pictóricos, basados en componentes principales y los desarrollados a partir de la obtención de distancias generalizadas.

- Métodos gráficos. Tal y como vimos en la figura 5.7 un diagrama de dispersión permite visualizar los casos correspondientes a dos variables de tal modo que los valores inusuales pueden ser detectados. La figura 5.8 muestra un ejemplo similar para dos

variables que corresponden a notas de alumnos en diversas asignaturas. Este gráfico presenta el inconveniente de que sólo es posible ver dos asignaturas. Para aumentar el número de asignaturas visibles simultáneamente una posibilidad es utilizar diagramas "girables" (spin-plots) que, al ser girados (Chambers, et al., 1983), permiten un cierto grado de visualización de la tercera dimensión a pesar de llevar a cabo esta representación en un lugar plano como es la pantalla del ordenador. En la figura 5.8 se muestran los datos acerca de un grupo de estudiantes en relación con sus notas en tres asignaturas. En el diagrama de dispersión de la izquierda podemos ver que la observación 105 parece destacar cuando se consideran las notas en Valenciano y Lengua. En la figura 5.9 vemos que esta puntuación sigue destacando cuando hacemos un diagrama en tres dimensiones y lo giramos de modo adecuado.

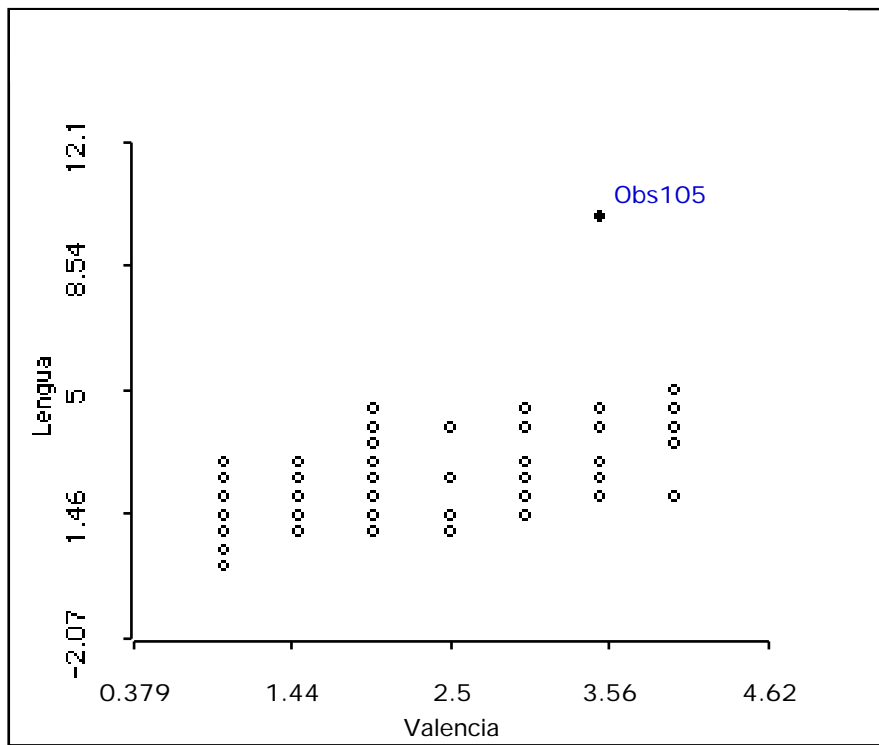


Figura 5.8: Diagrama de dispersión con outlier.

Esta estrategia sin embargo es de corto alcance. Aunque existen métodos para representar más variables simultáneamente (Velleman, 1995) en gráficos tipo 3D, en nuestra opinión, el usuario interesado debe realizar un entrenamiento especial antes de aspirar a percibir con claridad los gráficos implicados!

Una representación más sencilla de comprender y que en cierto modo combina la idea de una matriz de correlaciones con la de representación gráfica es la matriz de

diagramas de dispersión (Chambers, et al., 1983). En el gráfico de podemos ver un ejemplo.

la figura 5.10

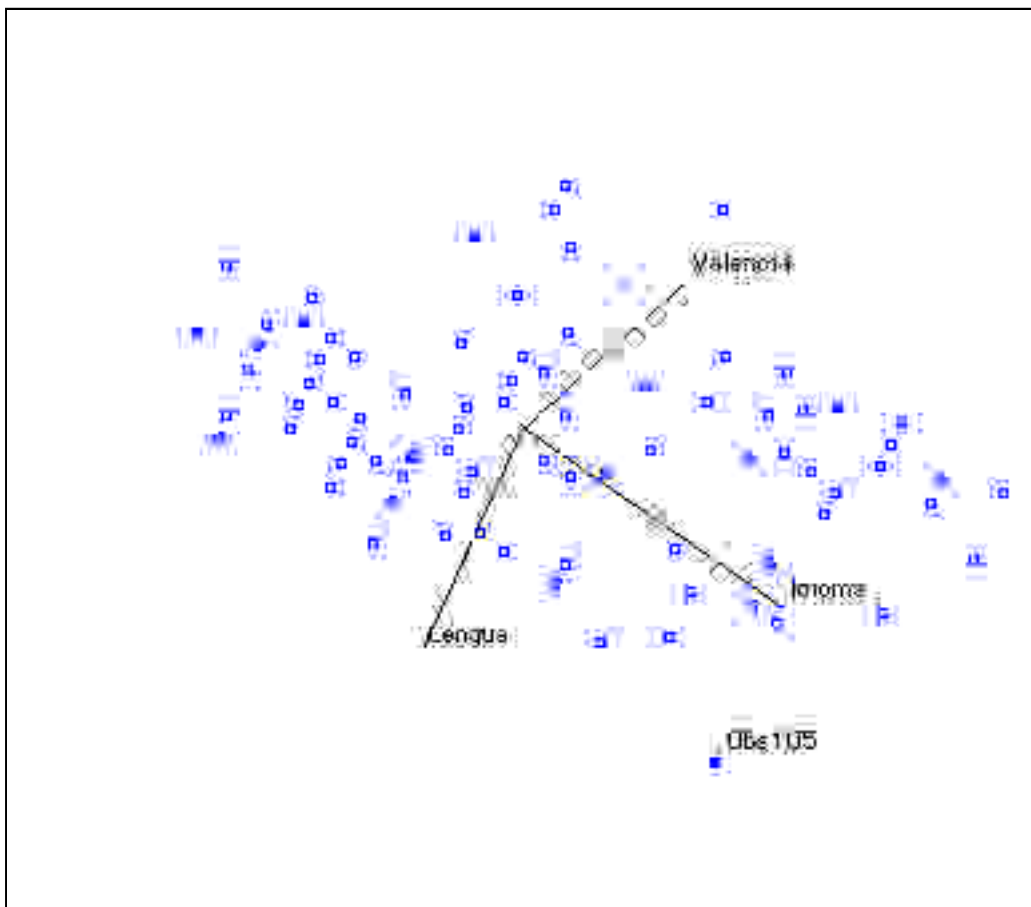


Figura 5.9 Gráfico 3D de dispersión con outlier.

La matriz de diagramas de dispersión permite incluir variables categóricas para de ese modo visualizar su cruce con las variables numéricas dando lugar a los conocidos por diagramas de puntos (*dotplots*), una representación a medio camino entre el diagrama de dispersión y el diagrama de cajas y bigotes. En la figura 5.11, se muestra una serie de categorías de fumadores junto a la edad para un grupo de sujetos. En él se puede ver a dos sujetos que muy probablemente correspondan a errores en los datos. Ambos a pesar de su corta edad han tenido tiempo tanto de convertirse en fumadores de cigarrillos como incluso de haber ya abandonado esta práctica!

- Métodos pictóricos: Los métodos pictóricos aplicados a las representaciones multivariadas consisten en general en asignar un símbolo a cada caso de tal manera que cada parte de este símbolo representen su valor en una variable. Chambers et al. (1983) describen varios de estos métodos tal y como estrellas, caras de Chernikov, perfiles,



arboles, etc. De entre ellos, hemos seleccionado como más apropiados los perfiles (figura 5.12 izquierda) y las estrellas (figura 5.12 derecha). Hemos representado las cinco variables del ejemplo anterior (notas en las asignaturas de Valenciano, Lengua, Idioma, Historia, Matemáticas y Ciencias Naturales representadas siguiendo este orden) para los 25 primeros sujetos. En los perfiles se indica el valor de la variable por medio de la altura, mientras que en las estrellas se indica mediante la distancia de los vértices respecto del centro. Puesto que todas las variables estaban en la misma escala no ha sido necesario transformarlas. En el lado izquierdo, perfiles "altos" indican buenas notas en general y "bajos" malas notas. En las estrellas, figuras grandes indican buenas notas y pequeñas malas. Quizás el individuo denominado g es el que destaca más rápidamente por sus grandes contrastes en una asignatura frente a las otras.

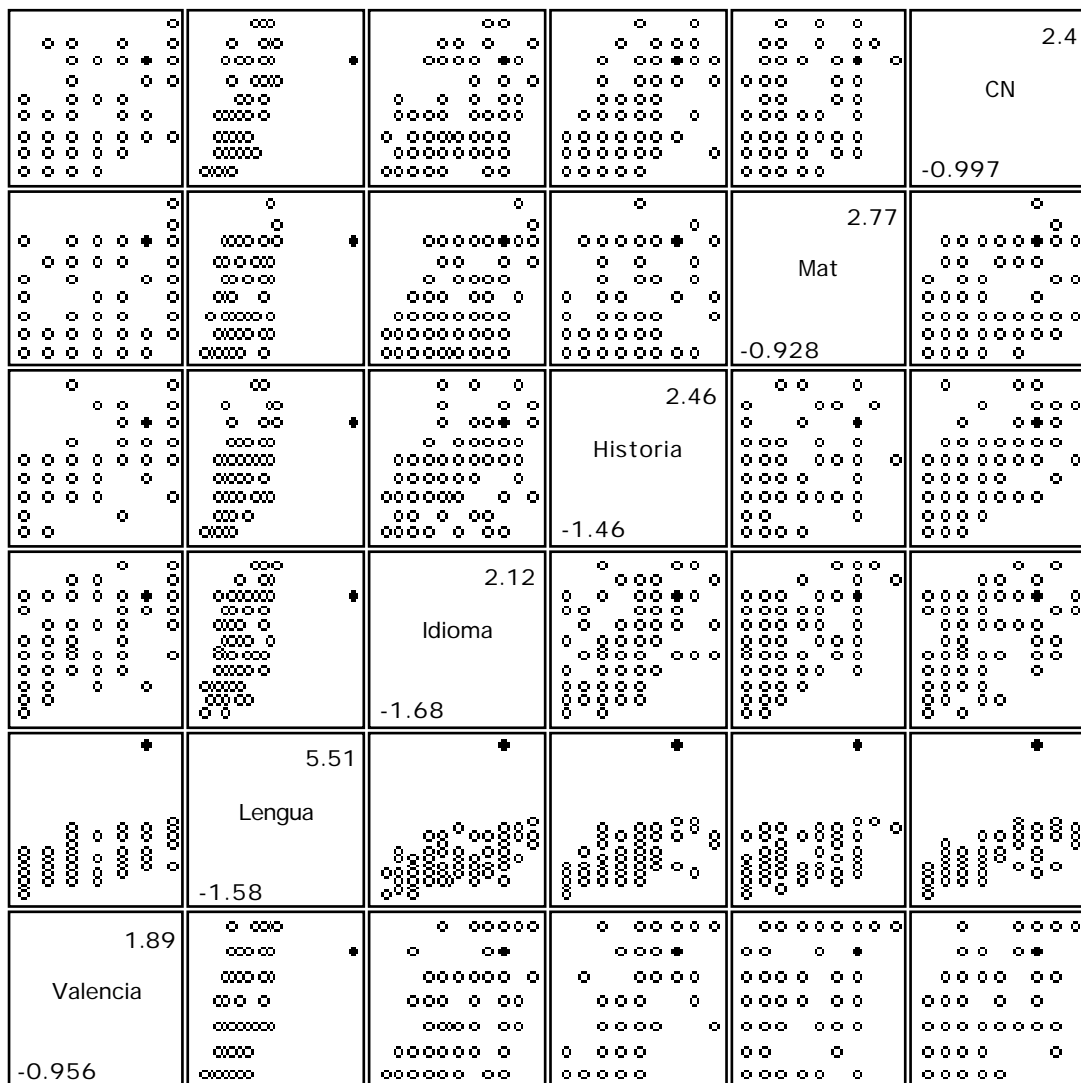


Figura 5.10. Utilización de la matriz de diagramas de dispersión para hallar outliers

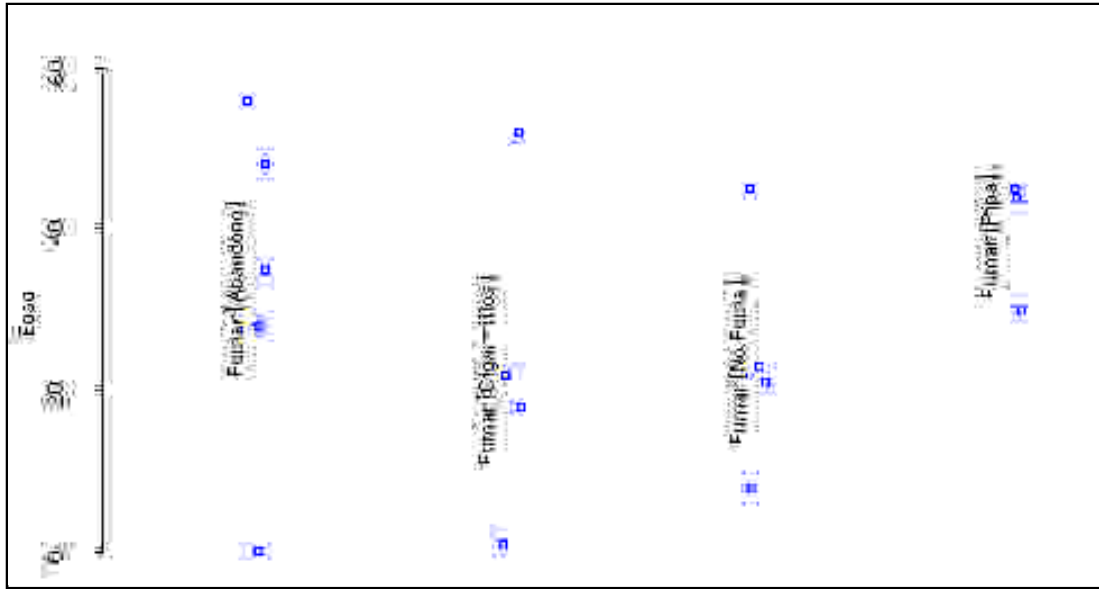


Figura 5.11. Diagrama de puntos para identificar errores cuando las variables implicadas son numéricas y categóricas

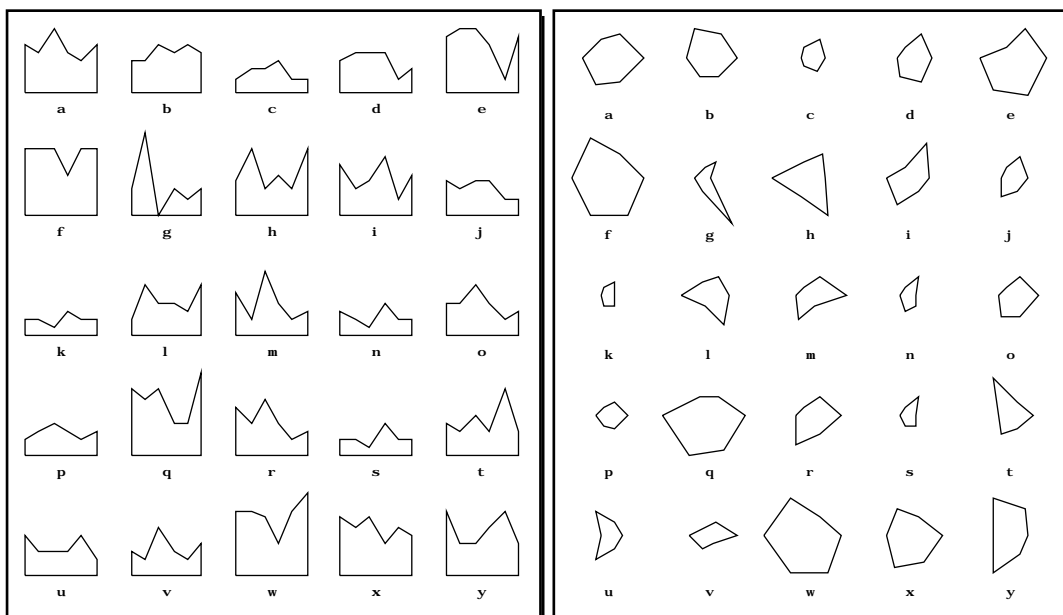


Figura 5.12. Gráficos pictóricos para detectar casos extraños.

• Métodos basados en componentes principales: Barnet y Lewis (1994) señalan que muchos autores han sugerido la idea de realizar un análisis de componentes principales sobre los datos, obtener las puntuaciones de los casos en éstos y representarlos gráficamente para examinar los valores extraños. Así, afirman que los primeros

componentes son sensibles a valores extraños hinchando las varianzas/covarianzas (o correlaciones, si los componentes principales han sido calculados utilizando correlaciones en lugar de covarianzas) mientras que los últimos serían sensibles a valores extraños que añadirán dimensiones espúreas u oscurecerán singularidades. Esto se muestra en los dos gráficos siguientes en los que se representan las puntuaciones en los componentes principales de los datos acerca de alumnos de escuela utilizando un bigráfico (Gabriel, 1986) lo cual añade información acerca de las variables. Podemos observar que la puntuación etiquetada como obs105 destaca enormemente en el primer componente principal (P0) probablemente haciéndolo coincidir casi perfectamente con la orientación de la variable Lengua al aumentar la varianza en esa variable. En el segundo gráfico vemos que en los dos últimos componentes principales hay otras puntuaciones que podríamos considerar sospechosas, aunque la obs105 sigue destacando enormemente.

• Distancias generalizadas: Barnet y Lewis consideran varias medidas de valores extremos del siguiente tipo:

$$R_i = (x_i - \bar{x})' S^p (x_i - \bar{x})$$

En donde  $R_i$  es un vector que indicaría la cercanía de un punto respecto del centroide de todos los datos y  $S$  correspondería con la matriz de varianzas/covarianzas. Diferentes valores de  $p$  darían lugar a diferentes medidas de distancia. Por ejemplo,  $p=0$  convertiría a  $R_i$  en distancias euclidianas al cuadrado. Más común es el caso con  $p=-1$  que da lugar a:

$$D_i = (x_i - \bar{x})' S^{-1} (x_i - \bar{x})$$

La cual es conocida como la distancia de Minkowsky (Aldenderfer and Blashfield, 1984) y tiene la ventaja de ponderar las distancias en función de las varianzas y las covarianzas con lo cual es posible descubrir observaciones que caen lejos del grupo de puntos general (Barnett y Lewis, 1994).

Un valor que suele ser discutido en relación con el análisis de regresión es el de *leverage* (influencia). Este valor es la diagonal de la matrix H en la siguiente expresión:

$$\hat{Y} = HY$$

Su relación con la distancia de Minkowsky es:

$$D_i \frac{1}{n-1} = H_{ii}$$

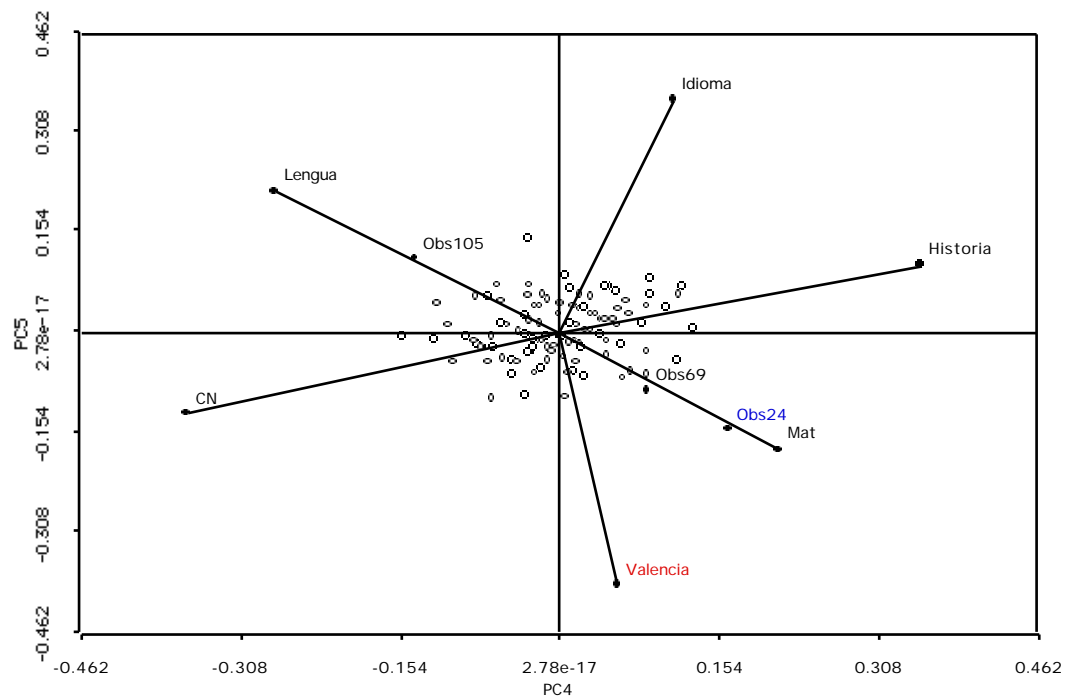
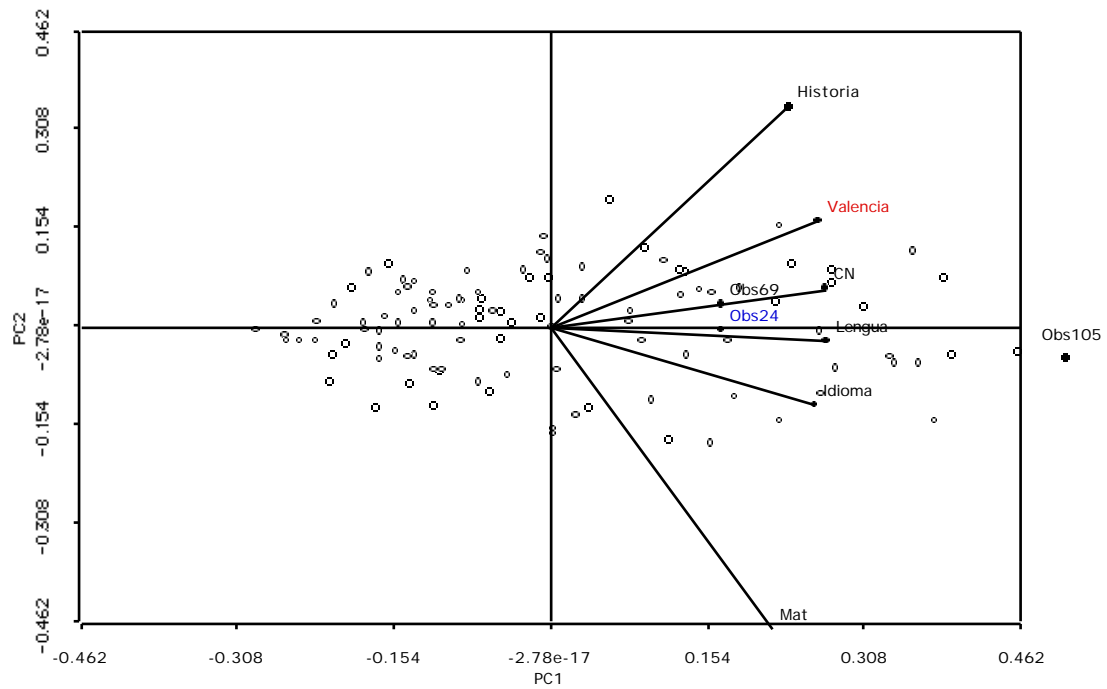


Figura 5.13. Biplots de componentes principales para detectar valores extraños.

La influencia tiene la ventaja de tener unos límites entre 0 y 1 por lo que puede resultar más fácilmente interpretable.

En cualquier caso, lo importante es que esta fórmula permite ordenar los casos en función de su lejanía con respecto a la masa central de los datos, del mismo modo que en el caso univariado podríamos realizar esta ordenación simplemente atendiendo a los valores originales de la variable. Barnett y Lewis discuten pruebas para determinar si un valor dado extremo puede considerarse que difiere del resto de una manera clara aunque dado nuestro objetivo de centrarnos solamente en errores resulta aconsejable simplemente examinar aquellos casos con puntuaciones en  $H_{ii}$  o  $D_i$  más altas.

En la figura 5.14 y 5.15 se muestran los componentes principales obtenidos en el análisis mostrado anteriormente junto con un gráfico de probabilidad normal de la influencia para las notas de estudiantes. Es posible ver que los sujetos con valores  $H_{ii}$  más altos son: el correspondiente a la puntuación 105 que ya hemos identificado como un *outlier*, y la 24 y la 69 que destacan mucho también en los dos últimos componentes principales.

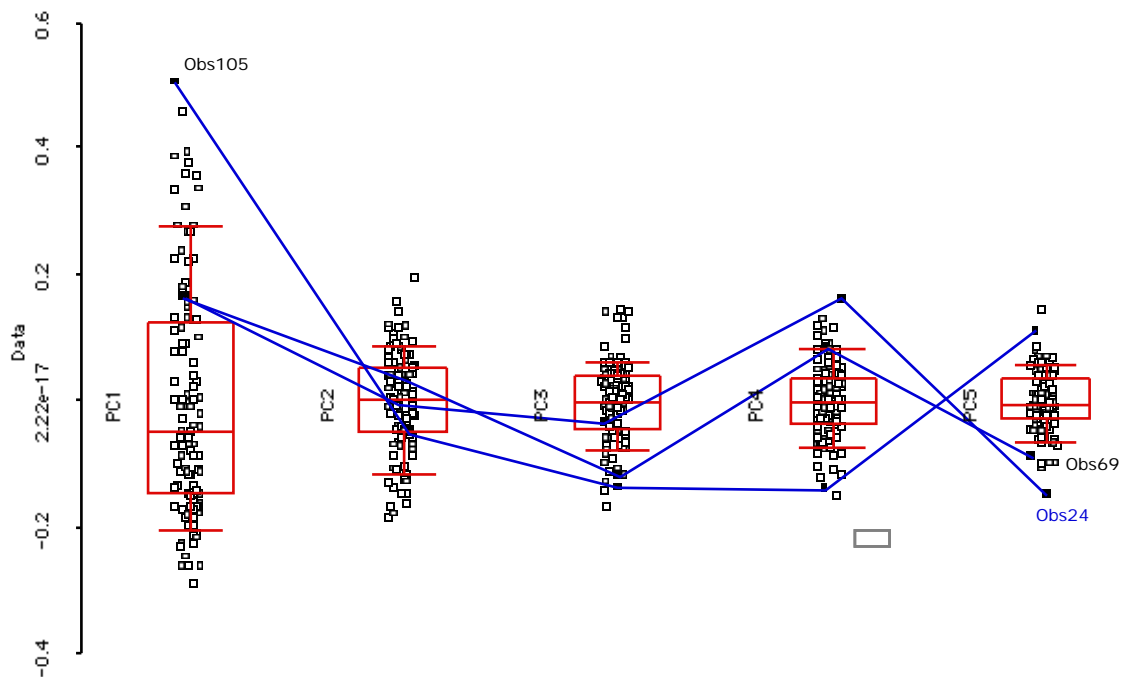


Figura 5.14. Boxplot de los componentes principales.

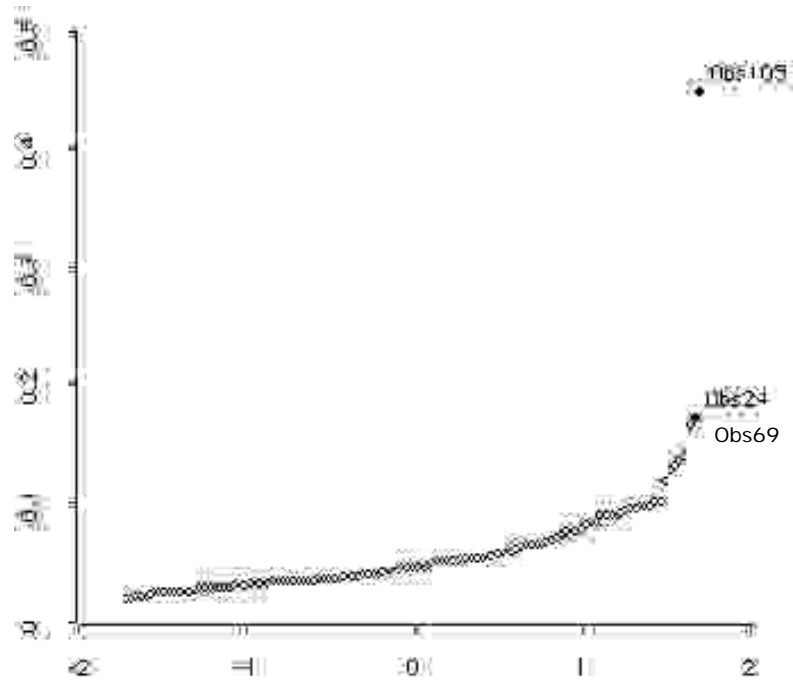


Figura 5.15. Gráfico de probabilidad normal de la influencia de los puntos







# 6

# *Datos Faltantes*

## 6.1. Introducción

El nombre de valores faltantes viene de la traducción del término inglés "missing", que, según el diccionario Webster (1983), puede definirse como ausente, faltante, no encontrado y, referente a personas, desaparecido. Aplicado a la situación del Proceso de Datos, un valor faltante es un dato del que carecemos y que, aunque sabemos que existe, no podemos averiguar. En el tipo de datos que suelen ser manejados en el contexto de la Psicología, estos datos faltantes suelen originarse en una *no-respuesta*: el sujeto, ante una serie de preguntas decide no contestar a alguna de ellas debido a que bien considera que vulneran de algún modo su intimidad, bien le resultan amenazadoras o, más simplemente, no conoce una respuesta adecuada.

Quizás el ejemplo más conocido de variable con tendencia a producir no-respuestas es la referida a los ingresos económicos del sujeto entrevistado. Muchas personas encuentran incómodo dar información acerca de sus salarios o patrimonio, y, después de haber contestado otras preguntas que consideran más irrelevantes, renuncian a responder acerca de estas cuestiones.

Una forma de evitar estos problemas es utilizar diseños de preguntas que evite ser demasiado directo. Por ejemplo, una estrategia utilizada para estimar el ingreso de los sujetos es preguntarles acerca de sus pertenencias o patrimonio tal y como el número de coches y/o modelos en su hogar, el tamaño de la casa en la que viven, los electrodomésticos, etc. Esto no obstante puede o no funcionar por lo que, en la práctica, el investigador se encontrará igualmente con una serie de datos que no le es posible averiguar, y que distorsionarán las inferencias que puede realizar.

Veremos a continuación una diferenciación entre la situación en que los datos están faltantes en parte o completamente. Seguiremos con una descripción de los mecanismos que dan lugar a valores faltantes, los cuales debemos suponer que subyacen a los datos observados. Posteriormente veremos los inconvenientes que presentan los análisis basados en datos completos o en datos disponibles, así como otras soluciones denominadas "rápidas". Por último, tenemos dos secciones que se refieren respectivamente a la exploración de las consecuencias de los tipos de borrados de datos faltantes y a la asignación-imputación de valores. Estas dos secciones constituyen la parte más avanzada del material aquí presentado.

## **6.2. Datos completa o parcialmente faltantes**

Bourque and Clark (1992) distinguen entre la situación en que no se conoce ninguna información acerca del dato faltante, y cuando se dispone de información parcial.

En general, podemos afirmar que el primer caso resulta poco probable, siempre y cuando el investigador tome un mínimo de precauciones. Por ejemplo, en una entrevista realizada en la calle siguiendo una serie de reglas que garanticen la aleatoriedad de los sujetos que son abordados, ciertos sujetos no contestarán. En ese caso, el entrevistador puede como mínimo registrar el sexo del sujeto que se ha negado a contestar y, aunque menos fiablemente, la edad aproximada. Estos datos, aun siendo mínimos suponen una información parcial que podría ser utilizada para estimar el impacto del género sobre las no-respuestas. Según Bourque and Clark (1992) es habitual que en las encuestas se consiga mayor cantidad de mujeres que de hombres debido a que aquellas suelen resistirse menos a contestar. Ello llevará a que éstas estarán *sobrerrepresentadas*, y los resultados globales podrán estar distorsionadas en caso de no tenerse en cuenta este fenómeno.

El ejemplo anterior es un recordatorio de los inconvenientes que poseen los diferentes métodos de recogida de respuestas. Por ejemplo, cuando se realizan las entrevistas en el domicilio de los sujetos es posible intentar acordar citas previamente, lo

cual disminuirá el número de sujetos que no contestan debido a situaciones pasajeras (falta de tiempo, otras actividades, etc.). También, es posible insistir en caso que se considere interesante. Finalmente, se puede obtener información indirectamente por medio de datos disponibles públicamente que ayuden a entender las causas del fenómeno.

Así pues, en la mayoría de los casos la información faltará sólo parcialmente. Ello nos permitirá realizar estimaciones del daño producido por la no-respuesta.

Una situación en la que sí que es posible obtener datos completamente faltantes es la de la recogida de datos mediante cuestionarios enviados por correo. Si estos envíos se producen de una manera "ciega", guiados únicamente por los datos del censo, la información acerca de los cuestionarios no devueltos será prácticamente nula.

### 6.3. Mecanismos que llevan a valores faltantes

Los mecanismos que dan lugar a valores faltantes son de gran importancia porque de ellos se derivan las consecuencias que se producirán sobre los datos completos. Hay tres criterios posibles de clasificación de los mecanismos:

a) En función de que la probabilidad de respuesta de una variable Y, dadas dos variables X e Y, dependa de alguno de los

(1) sea independiente de Y y de X, lo cual llamaremos Datos Faltantes Completamente al Azar (FCA),

(2) dependa de X pero no de Y, lo cual llamaremos Datos Faltantes al Azar (FA), y,

(3) dependa de Y y posiblemente también de X, caso que implica que los datos no faltan al azar y no son ni FCA ni FA.

b) En función de que el mecanismo que lleva a producir los valores faltantes sea conocido o no. Así tenemos:

(1) mecanismos conocidos: en ocasiones se sabe cuál es la forma o razón que produce la existencia de valores faltantes. A veces ese mecanismo está bajo el control del investigador, como, por ejemplo, cuando se lleva a cabo un *doble muestreo*. En este caso se selecciona una muestra grande de sujetos y se registra una serie de variables, para, posteriormente, seleccionar una submuestra de la muestra original y medirse otras variables, quizás más caras o más costosas de conseguir. Incluso, cuando se lleva a cabo un *muestreo aleatorio simple* de una

población puede pensarse (aunque no es lo usual) en que los sujetos no incluidos son valores faltantes. Un ejemplo de un mecanismo que es conocido pero no está bajo el control del investigador es cuando se produce un *censurado* de los datos, y sólo se registran valores dentro de ciertos límites (por ejemplo, en estudios de seguimiento de casos en pacientes crónicos, algunos de ellos pueden fallecer durante la realización del experimento, por lo que tendrán menos mediciones que el resto).

(2) mecanismos desconocidos: cuando ciertas mediciones no pueden ser obtenidas pero no se conoce el mecanismo subyacente el investigador se encuentra ante la situación de tratar con un problema que potencialmente puede distorsionar sus resultados. Por ejemplo, una pregunta sobre ingresos no es contestada por ciertos sujetos. En ausencia de otra información complementaria nos encontraríamos ante valores faltantes que siguen un mecanismo desconocido. Cuando como es probable, los valores faltantes dependan de la variable considerada (cuando por ejemplo los salarios no contestados sean los más altos) pero no se conozca exactamente el mecanismo nos encontraremos con problemas de inferencia.

c) en función de su *ignorabilidad*: la ignorabilidad del mecanismo que lleva a datos faltantes es importante para salvaguardar las inferencias que podemos extraer de ellos. Según Little y Rubin (1987), el mecanismo de datos faltantes es ignorable cuando es FCA en todos los casos, y, siempre y cuando se utilicen métodos de estimación especiales, cuando es FA. El mecanismo no es ignorable cuando los datos no sean ni FCA ni FA. Veamos con más detalle las razones que subyacen a esta afirmación.

(1) El mecanismo de datos faltantes es ignorable bajo FCA: Siguiendo el ejemplo de Little y Rubin (1987), si tenemos  $X$ =edad e  $Y$ =salario, la no-respuesta sólo afecta a  $Y$ , y la probabilidad de que se responda el salario es independiente tanto de la edad como del salario, entonces los datos son FCA y la muestra de sujetos que contestan puede considerarse una muestra aleatoria del total de los sujetos. En este caso los sujetos que no contestan se supone que no forman un subconjunto del total especial por alguna razón, y, si el investigador logra justificarlo adecuadamente, sus inferencias pueden continuar de una manera normal excepto por el hecho que se producirá una reducción con respecto a la muestra que originalmente tenía planeado recoger. De todos modos, si se cumplen estas condiciones, el investigador podría plantearse seguir su búsqueda hasta obtener el tamaño muestral originalmente planeado.

(2) Un caso mucho más interesante que el anterior es cuando los datos son FA. Para explicar este caso, supongamos que partimos de la situación en que la variable Y está sesgada por las no-respuestas y sólo se ha recogido una submuestra del total, por lo que las estimaciones de la media y la desviación típica de esa variable serían incorrectas. En el caso del ingreso, supongamos que sólo se han recogido los ingresos medios y bajos. Sin embargo, examinando la relación entre ambas descubrimos que la variable X está relacionada con Y y que, como veremos más adelante, podemos utilizar esta información para mejorar la estimación de la media de Y. Supongamos que los sujetos con edad media son los de valores más altos en salario. Podemos asignarles un valor medio correspondiente a los sujetos con esa edad que sí contestaron a los que no contestaron, incrementando de este modo el número de sujetos con salarios altos. Las estimaciones de la media de Y serán ahora más correctas. De este modo, aunque los datos faltantes en Y dependen de los valores de ella misma, al haber sido estimados mediante X, podemos asumir que esa dependencia ha desaparecido y el sesgo en la variable Y ha disminuido. Naturalmente, si en lugar de una variable X tenemos varias la reducción en el sesgo de la variable Y puede ser considerablemente mayor. FA puede ser entendido como un caso faltante causado por una variable que ha sido medida e incluida en el análisis (Graham, et al., 1996). Ello nos permitiría ignorar el mecanismo produciendo casos faltantes. Una situación especialmente interesante es cuando los datos son faltantes de modo multivariado y es necesario realizar estimaciones en las que todas ocupan de modo sucesivo el papel de la variable Y. Veremos más adelante métodos para tratar esta situación.

La situación anterior puede verse complicada si, examinando los ingresos en grupos condicionados con respecto a la edad, observáramos que existe una tendencia a que el ingreso no sea registrado para aquellos que tienen mayores ingresos nos encontraríamos ante una situación en que los valores faltantes no son ni FCA ni FA. Este mecanismo de datos es no ignorable y resulta muy difícil de tratar mediante métodos puramente analíticos.

## 6.4. Consecuencias de los valores faltantes

Las consecuencias de los valores faltantes dependen de la complejidad de los datos que estemos manejando y del tipo de preguntas que estemos interesados en contestar. Distinguiremos entre la situación en la que sólo disponemos de una variable y cuando, más habitualmente, tenemos varias variables. Veamos cada uno de estos casos.

a) Sólo una variable.

Supongamos que tenemos una población de tamaño  $N$  de la que es seleccionada una muestra de tamaño  $n$  para registrarse una variable  $Y$ ,  $y$ , de esos  $n$  casos, sólo  $m$  responden. La diferencia  $n-m$  son por tanto casos faltantes. Si podemos asumir que nuestros datos son FCA entonces (Little and Rubin, 1987) un intervalo de confianza del 95% para la media poblacional será:

$$\bar{y}_R \pm 1.96 \sqrt{(s_R^2/m) - (s_R^2/N)}$$

En donde  $\bar{y}_R$  y  $s_R^2$  son la media y la varianza de las unidades que responden. Asumiendo un tamaño de la población infinito esta fórmula se convierte en la habitual.

El punto clave aquí es que podamos asumir que nuestros datos son FCA ya que en caso contrario, Cochran (1963) muestra que  $\bar{y}_R$  es un estimador sesgado de  $\bar{Y}$ , con un sesgo aproximado de:

$$\text{sesgo}(\bar{y}_R) = \bar{Y}_R - \bar{Y} = (1 - r)(\bar{Y}_R - \bar{Y}_{NR})$$

En donde:

$\bar{Y}$  = Media de la población

$\bar{Y}_R$  = Media de los que responden

$r$  = Proporción de unidades que responden en la población.

$\bar{Y}_{NR}$  = Media de los que no responden

Naturalmente, si asumimos FCA,  $\bar{Y}_{NR}$  y  $\bar{Y}_R$  son iguales y el sesgo es nulo. En caso contrario, al no tener información acerca de  $\bar{Y}_{NR}$  no podemos corregir ese sesgo. La solución al problema vendrá pues de la obtención de otras fuentes de información que nos permitan estimar  $\bar{Y}_{NR}$  de modo indirecto, aumentando por tanto el número de variables consideradas en nuestros estudios.

De todos modos, aun asumiendo FCA, dado que el número  $m$  de respondientes es menor que el tamaño de la muestra,  $n$ , originalmente planeado recoger nos encontramos ante un intervalo de confianza más grande que el previsto, con los consiguientes problemas de precisión derivados.

b) Varias variables recogidas simultáneamente.

Podemos diferenciar dos problemas:

1) Pérdida de la muestra debido a la existencia de valores faltantes en diferentes variables. Puesto que muchos métodos de análisis y también el software disponible está diseñado para ser utilizado solamente con casos completos, todos los que incluyan un dato faltante no podrán ser utilizados. Esto produce como resultado una pérdida enorme de datos. Por ejemplo, asumiendo un número de variables  $K=20$ , en donde cada variable es observada independientemente siguiendo un proceso de Bernoulli con un 5% de valores faltantes, entonces la proporción esperada de datos faltantes puede ser calculada mediante la fórmula de la distribución binomial.

$$p(x) = \frac{N!}{X!(N - X)!} p^x q^{N-x}$$

Para el caso en que  $p(x=0)$  la probabilidad es:  $0.95^{20}=0.35$ . Es decir, aproximadamente un 35% de los casos *no* tendrían valores faltantes. Este 35% es el que quedaría en caso de querer llevarse a cabo un borrado *listwise* (el método por defecto en la mayoría de los paquetes estadísticos).

Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>
1	1	•
2	2	•
3	3	•
4	4	•
1	•	1
2	•	2
3	•	3
4	•	4
•	1	4
•	2	3
•	3	2
•	4	1

*Tabla 6.1. Inconsistencia de las correlaciones pairwise.*

2) Cambio de base de variable a variable en función de las diferentes bases muestrales. La situación anterior se producía por la eliminación de todos los casos que tuvieran algún valor faltante. Una alternativa natural es no eliminar el caso entero sino utilizar los valores que hayan sido obtenido en las variables respectivas. Este enfoque presenta el inconveniente que, al cambiar las bases por las que se calculan estadísticos

sencillos como la media o la proporción muestral, muchas comprobaciones sencillas acerca de la corrección de resultados dejan de poderse hacer.

Una situación muy común producto de la indefinición de las bases es lo que ocurre al calcular matrices de correlaciones entre variables con diferentes proporciones de valores faltantes. En esas matrices, cada valor responde a un número diferente de sujetos con lo que se pueden producir resultados indeterminados tal y como el ejemplo artificial de la tabla 6.1 puede ayudar a ilustrar (Little and Rubin, 1987). En esos datos,  $r_{y_1 y_2} = 1$  y  $r_{y_1 y_3} = 1$ . Esto implicaría  $r_{y_2 y_3} = 1$ . Sin embargo, el resultado es  $r_{y_2 y_3} = -1$ .

## 6.5. Análisis de casos completos frente a análisis de datos disponibles

Empezaremos definiendo estos dos términos:

- **Datos Completos:** Nos referiremos a datos completos cuando sólo consideramos en los análisis aquellos casos con todos los datos completos en las variables consideradas. Este tratamiento de los datos faltantes es llamado habitualmente en los paquetes estadísticos como *Listwise*, nombre que se refiere a que la eliminación de un caso se produce para toda la fila cuando existe un dato faltante. Este procedimiento presenta el inconveniente de eliminar mucha de la información disponible, corriéndose además el riesgo de que las variables faltantes no sean FCA o FA y se produzcan problemas de muestreo. No obstante, muchos procedimientos de análisis necesitan datos completos para ser ejecutados por lo que éste procedimiento es aplicado automáticamente por los paquetes estadísticos. Más importante, transformaciones sencillas, tal y como la suma de variables, no pueden ser calculadas para un caso cuando uno de sus valores es faltante.

- **Datos Disponibles:** Hablaremos de datos disponibles cuando para hacer un análisis tengamos en cuenta todos los casos disponibles para esa variable. En la sección siguiente discutiremos algunas de las fórmulas posibles para el cálculo de las matrices de varianzas y covarianzas utilizando los datos disponibles de variables. Antes de ello, no obstante, discutiremos uno de las situaciones más importantes, el cálculo de matrices de correlaciones utilizando el método *Pairwise* v. *Listwise*



### 6.5.1. Datos completos frente a datos disponibles

Un caso muy importante de utilización de datos disponibles es el que aparece al calcular correlaciones entre pares de variables. En él un caso es utilizado cuando ninguno de sus valores para las dos variables implicadas es faltante. Este tratamiento de los datos faltantes se denomina *Pairwise* y presenta el inconveniente que, cuando se calcula una matriz de intercorrelaciones entre varias variables, el número de casos disponibles variará en función del par considerado. Esta variación produce una serie de problemas que describiremos a continuación por medio de un ejemplo tomado de Schafer (1997). En él se muestra unos datos provenientes de Raymond (1983) acerca de un estudio sobre actitudes en relación con el aprendizaje de una lengua extranjera para hablantes ingleses. Estos datos incluyen variables indicadas en la tabla 6.2:

LAN: Lenguaje Extranjero estudiado (1=Francés; 2=Español; 3=Alemán; 4=Ruso)
EDAD: Grupo de edad (1=menos que 20; 2=20-21; 3=22-23; 4=24-25; 5=26 o más)
PREV: Número de cursos de lenguaje extranjero previos (1=ninguno; 2=1, 3=2, 4=3, 5=4 o más) <sup>2</sup>
SEXO: 1=Hombre, 2=Mujer
EALE: Escala de actitudes hacia los lenguajes extranjeros.
TALM: Test Moderno de Aptitudes para el Lenguaje.
TAE-V: Test de Aptitud Escolar (Verbal).
TAE-M: Test de Aptitud Escolar (Matemático).
ING: Puntuación en Inglés en el examen de ingreso en la universidad.
CALI: Calificaciones promedio en el instituto.
CALA: Calificaciones actuales en la universidad.
NLEX: Nota final en Lengua Extranjera. (4=A; 3=B; 2=C; 1=D; 0=F).

Tabla 6.2. Variables en el ejemplo de Raymond (1983)

En el gráfico 6.3 y en la tabla 6.4 es posible ver los valores faltantes por variable junto con algunos estadísticos descriptivos. Estas tablas y gráficos corresponderían a los datos *disponibles*.

<sup>2</sup> Aunque no corresponde a este tema, aquí tenemos un buen ejemplo de una mala codificación de los datos.

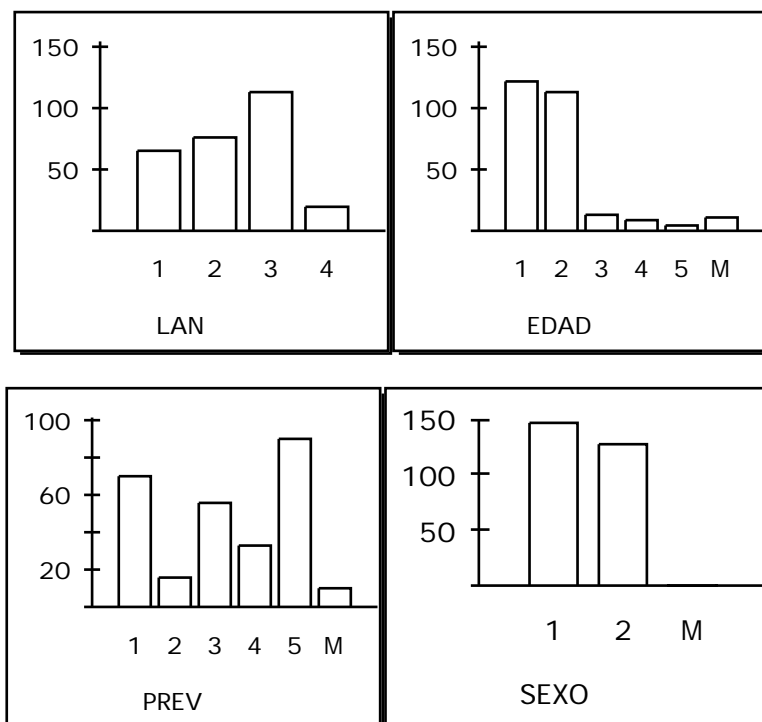


Figura 6.3. Diagramas de barras para las variables categóricas en el ejemplo enseñanza de idiomas.

Podemos observar que el número de valores faltantes por variable va desde cero hasta el máximo de 49 (variable TALM), que supone un 17% del total de casos. Sumando el número de valores faltantes (259) y dividiendo por el total de valores de la tabla multivariada ( $279 \times 12 = 3348$ ) podemos obtener la proporción de valores faltantes en la tabla. Este valor (en porcentajes) es 7.73%.

La tabla 6.6 muestra los estadísticos descriptivos para los casos *completos*. En ella observamos en primer lugar que la columna de Faltantes tiene siempre el mismo valor (105). Esto supone un  $(105/279) \times 100 = 37.6\%$  del total de casos. Esto significa que la utilización de técnicas multivariadas que incluyeran todas estas variables por medio de las rutinas utilizadas habitualmente en muchos paquetes estadísticos podría suponer una reducción del número de casos que podríamos considerar como desmesurado.

Variable	Faltant.	Media	Varianza	DvTípica	Min	Max
LAN	0	-	-	-	1	4
EDAD	11	-	-	-	1	5
PREV	11	-	-	-	1	5
SEXO	1	-	-	-	1	2
EALE	0	82.4373	197.218	14.0434	28	110
TALM	49	24.3304	39.7681	6.30619	9	40
TAE-V	34	504.331	8195.27	90.5277	210	790
TAE-M	34	564.249	8248.43	90.8209	180	800
ING	37	53.9256	241.140	15.5287	8	113
CALI	1	2.76741	0.365177	0.604299	1.2	4
CALA	34	3.29571	0.226819	0.476255	2	4
NING	47	3.32759	0.758024	0.870645	0	4
259						

Total de casos=279

Tabla 6.4.Descriptivos para el ejemplo de enseñanza de idiomas.

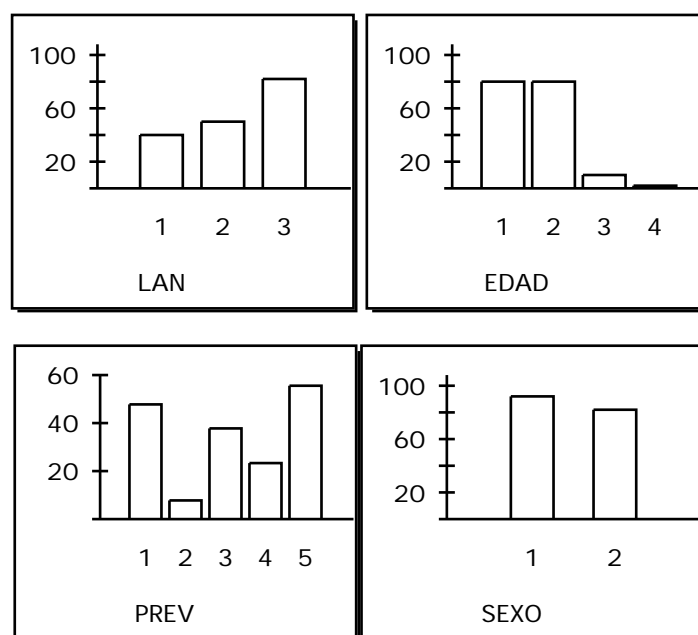


Figura 6.5.Diagramas de barras para los datos completos en las variables categóricas en el ejemplo de enseñanza de idiomas

Esta reducción afecta a muchos aspectos del análisis de datos aparte del ya citado de disminución de la muestra. Entre los problemas más interesantes podemos destacar que se ha producido una reducción en los valores de las variables categóricas. Así, todos los

estudiantes de Ruso (categoría 4 en variable LAN) ya no se encuentran disponibles<sup>3</sup>, así como los que se encuentran en la categoría 5 de EDAD. El resto de las variables también ha visto afectados sus respectivos estadísticos descriptivos ligeramente tal y como el lector puede examinar por sí mismo en los gráficos de la figura 6.5 y en la tabla 6.6.

Variable	Faltant.	Media	Varianza	DesvTíp.	Min	Max
LAN	105	-	-	-	1	3
EDAD	105	-	-	-	1	4
PREV	105	-	-	-	1	5
SEXO	105	1.47126	0.250615	0.500614	1	2
EALE	105	83.0690	160.689	12.6763	44	110
TALM	105	25.4023	37.8835	6.15495	9	40
TAE-V	105	511.328	7742.28	87.9902	300	790
TAE-M	105	575.063	7965.09	89.2473	180	800
ING	105	55.0057	242.861	15.5840	19	113
CALI	105	2.86011	0.347706	0.589666	1.3	4
CALA	105	3.36339	0.209493	0.457704	2.2	4
NING	105	3.40230	0.727394	0.852874	0	4

Tabla 6.6: Descriptivos para los datos completos en el ejemplo de enseñanza de idiomas.

	SEXO	EALE	TALM	TAE-V	TAE-M	ING	CALI	CALA	NING
SEXO	1.000								
EALE	0.278	1.000							
TALM	0.186	0.159	1.000						
TAE-V	-0.030	0.009	0.417	1.000					
TAE-M	-0.195	-0.074	0.433	0.392	1.000				
ING	0.087	0.131	0.564	0.677	0.412	1.000			
CALI	0.060	0.051	0.413	0.230	0.274	0.306	1.000		
CALA	0.234	0.047	0.498	0.284	0.315	0.321	0.432	1.000	
NING	0.172	0.248	0.433	0.071	0.124	0.212	0.524	0.411	1.000

Tabla 6.7. Matriz de correlaciones obtenida con borrado Pairwise

La misma idea de solamente considerar los casos disponibles para realizar el cálculo actual puede ser extendida a análisis por pares de variables, tal y como los necesarios para conseguir matrices de intercorrelaciones. Estas matrices son muy importantes por que son el punto de partida del cálculo de muchas técnicas multivariadas. En la tabla 6.7 se muestran las correlaciones pairwise en la 6.8 el número de casos disponible para el

<sup>3</sup> A pesar que originalmente no había ningún valor faltante en esa variable.

cálculo de ellas. Podemos ver que el número de casos disponible varía por par de variables, en ocasiones de manera muy marcada.

	SEXO	EALE	TALM	TAE-V	TAE-M	ING	CALI	CALA	NING
SEXO	278								
EALE	278	279							
TALM	229	230	230						
TAE-V	244	245	200	245					
TAE-M	244	245	200	245	245				
ING	241	242	197	242	242	242			
CALI	277	278	229	245	245	242	278		
CALA	244	245	200	245	245	242	245	245	
NING	232	232	201	207	207	206	231	207	232

*Tabla 6.8. Matriz de casos considerados para el cálculo de matrices de correlaciones pairwise*

El gráfico de la figura 6.10 representa las diferencias entre las correlaciones utilizando un método u otro. En este caso las diferencias no son muy grandes aunque en otros lo pueden ser. Como es posible ver las diferencias no son muy grandes y alcanzan como máximo un valor de 0.075 (tanto en signo positivo como en signo negativo).

	SEXO	EALE	TALM	TAE-V	TAE-M	ING	CALI	CALA	NING
SEXO	1.000								
EALE	0.323	1.000							
TALM	0.186	0.105	1.000						
TAE-V	0.001	0.043	0.425	1.000					
TAE-M	-0.181	-0.138	0.394	0.319	1.000				
ING	0.122	0.181	0.579	0.721	0.396	1.000			
CALI	0.067	0.057	0.375	0.227	0.204	0.318	1.000		
CALA	0.227	0.057	0.513	0.279	0.264	0.332	0.404	1.000	
NING	0.136	0.225	0.391	0.140	0.171	0.254	0.516	0.452	1.000

*Tabla 6.9. Matriz de correlaciones obtenida con borrado Listwise*

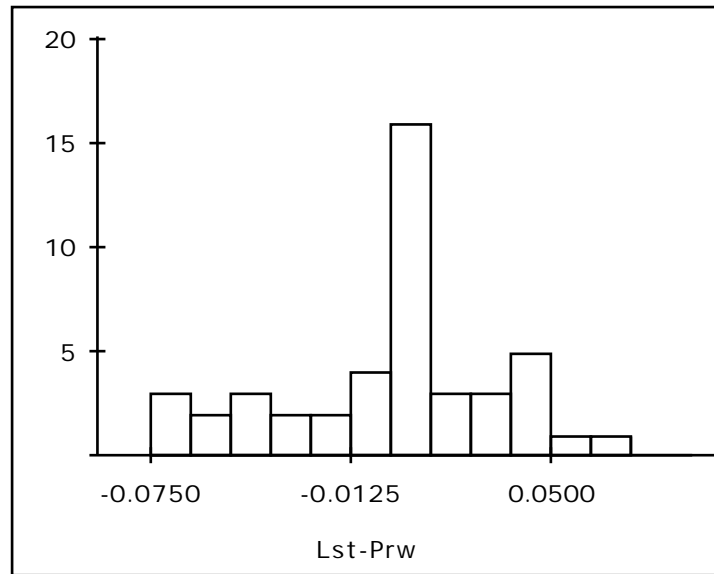


Figura 6.10: Histograma de diferencias entre correlaciones calculadas mediante Listwise y Pairwise.

La mayoría de los análisis multivariados están diseñados para utilizar sólo datos completos, por lo que realizarán como primer paso una eliminación *listwise*. Hemos visto que esto presenta el inconveniente de la eliminación de un gran número de casos. Una alternativa que podría plantearse es utilizar versiones de análisis multivariados que acepten matrices de correlaciones (o de varianzas-covarianzas) como punto de partida en lugar de los datos originales y utilizar para ellos las matrices *pairwise*. Esto en apariencia presentaría la ventaja de utilizar un número mayor de casos para el análisis con lo que nuestros resultados parecerían más correctos. No obstante, esta no es una buena idea por las siguientes razones:

1) Puesto que las correlaciones han sido calculadas mediante diferentes tamaños muestrales, y el error típico de la distribución muestral de  $r$  está basado en esos tamaños, nos encontraremos con que cada correlación tiene una estabilidad diferente (Tabachnik y Fidell, 1989). Si la interpretación que se hiciera de ellas fuera individual esto quizás podría aceptarse, pero su utilización conjunta en pruebas de hipótesis (cuando se calcula un análisis de regresión por ejemplo) es totalmente cuestionable puesto que no existe ningún conocimiento acerca de pruebas de hipótesis en esta clase de situaciones (Norusis, 1990).

2) La matriz de correlaciones calculada mediante el método *Pairwise* no es consistente. El ejemplo artificial mostrado en la sección anterior ilustra como la correlación entre dos variables era cero cuando en realidad debería ser uno al utilizar el

método *Pairwise*. En concreto, si tenemos una matriz de correlaciones de 3 variables la siguiente relación debería cumplirse:

$$r_{13}r_{23} - \sqrt{(1 - r_{13})(1 - r_{23})} < r_{12} < r_{13}r_{23} + \sqrt{(1 - r_{13})(1 - r_{23})}$$

No obstante, cuando utilizamos correlaciones *Pairwise*  $r_{12}$  puede salir de rango.

3) Una última consecuencia está relacionada con los autovalores de una matriz. En principio, los autovalores de una matriz simétrica (Amon, 1991; Green, 1993) son reales, lo cual es condición para que una matriz pueda considerarse semipositiva definida. En el caso de una matriz de varianzas-covarianzas esto es intuitivamente comprensible puesto que los autovalores no son más que una consolidación de la varianza incluida en la diagonal de esa matriz, y una varianza, por definición, no puede ser negativa.

No obstante, una matriz *pairwise* puede producir autovalores negativos (Little and Rubin, 1987; Tabachnik and Fidell, 1989). Estos autovalores, por tanto, estarían haciendo referencia a varianzas negativas, lo cual resulta difícil de asimilar. Además, si pensamos en la varianza total como en un valor fijado (tal y como en una matriz de correlaciones en donde este valor es igual a la suma de la diagonal), la existencia de eigenvalores negativos producirá una compensación que llevará a hinchar los valores positivos. Los resultados obtenidos bajo estas condiciones estarán por tanto distorsionados.

### 6.5.2. Otras Soluciones al problema de los valores faltantes utilizando la matriz de datos disponibles

Siguiendo con la idea de utilizar los datos disponibles para el cálculo de los estadísticos necesarios para análisis multivariados Little y Rubin (1987) describen una serie de fórmulas alternativas que intentan aprovechar todavía más la información disponible. Así tenemos:

$$r_{jk} = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}$$

En donde el  $s_{jk}^{(jk)}$  hace referencia a la covarianza calculada de modo *pairwise* y  $s_{jj}^{(j)}$  y  $s_{kk}^{(k)}$  hace referencia a la desviación típica para todos los casos disponibles indicados en

el sufijo. Esta fórmula presenta el inconveniente que las estimaciones de las correlaciones pueden superar los límites de -1 o +1.

Otra posibilidad es estimar la covarianza de la fórmula anterior de este modo:

$$s_{jk}^{(jk)} = \frac{1}{(n^{(jk)} - 1)} \sum_{(jk)} (y_{ij} - \bar{y}_j^{(j)})(y_{ik} - \bar{y}_k^{(k)})$$

Esta fórmula aprovecha todos los casos disponibles para el cálculo de las medias. Combinada con la fórmula anterior corresponde al método ALLVALUE disponible en BMDP. En general, estas dos fórmulas tienden a ser no positivas semidefinidas más a menudo que el método descrito anteriormente (Dixon, 1981). Little y Rubin (1987) señalan que, aunque los métodos basados en los casos disponibles puedan parecer en apariencia superiores debido a que aprovechan más información de la muestra, esto sólo parece cumplirse cuando los datos son FCA y las correlaciones son moderadas. Cuando las correlaciones son altas los métodos basados en análisis de datos completos parecen ser superiores. En cualquier caso, sin embargo, ningún método parece como satisfactorio en general.

## 6.6. Exploración de valores faltantes.

En secciones anteriores hemos expuesto el tema de la depuración de datos. En él describíamos métodos para detectar irregularidades en los datos susceptibles de necesitar corrección. Generalmente, las tareas que describiremos a continuación se desarrollarán simultáneamente a las descritas en el apartado de depuración de datos. No obstante, la revisión de los datos faltantes incorpora problemas que merecen ser tratados por separado de los de la depuración de datos.

Aunque no es posible realizar una lista exhaustiva de los posibles problemas, ya que cada situación puede generar los suyos particulares, sirva la siguiente como ejemplo de las posibles:

- Entrevistadores más propensos a utilizar la/s categoría/s de datos faltantes.
- Zonas geográficas o momentos temporales más propensos a generar datos faltantes.
- Lugar de realización de la entrevista (lugar de trabajo, hogar, etc.).



- Individuos con ciertas actitudes sociales no contestan a preguntas acerca de sus ingresos.
- Cuando hay una relación de necesidad entre tener un valor en una variable o no tenerlo (por ejemplo, estar matriculado en un curso y tener nota en sus diferentes exámenes).
- Preguntas íntimas.
- Ingresos.
- Conductas delictivas o susceptibles de generar unas consecuencias negativas para el entrevistado a pesar que se le garantice el anonimato.

Existen dos problemas en relación con la exploración de los datos faltantes que es interesante comentar:

a) A menudo los paquetes estadísticos no los incorporan dentro de las técnicas estadísticas o gráficas. Ello hace que su existencia no sea tenida en cuenta por el analista y, por tanto, pierda la pista de sus posibles efectos. Así, sólo existe un programa, especializado precisamente en representaciones gráficas para el análisis de valores faltantes, que incorpora histogramas o diagramas de dispersión que incluyen información acerca de aquellos y del que mostraremos algunas de sus capacidades en esta sección.

b) La existencia de dos cuestiones que es necesario explorar simultáneamente. Estas son, en primer lugar, la relación entre los valores faltantes para las variables, y, en segundo lugar, la conexión entre esos valores faltantes y los valores presentes. Este problema puede representarse esquemáticamente con la figura 6.11.

En el lado izquierdo tenemos una matriz de datos normal (se indica con asteriscos los valores faltantes). En el lado derecho tenemos esa matriz transformada en una dicotomizada, en la que si un valor está presente se indica con un 1 y si falta con un 0. Cuando estamos analizando los valores faltantes una pregunta que nos haremos es la referida a las relaciones de tipo A (entre las variables dicotomizadas), y las de tipo B, entre las variables dicotomizadas y las variables normales. Naturalmente, nuestro fin último es examinar las relaciones de tipo C. En general, los problemas surgen al examinar las relaciones de tipo B, pues la existencia de valores faltantes dificulta ciertos cálculos. Por ejemplo, la relación entre la variable  $D_4$  y la variable  $X_2$  no podría ser obtenida debido a la existencia de estos faltantes. Algunas técnicas o métodos de los mostrados a continuación son capaces de mostrar ambos las relaciones de tipo A y B simultáneamente (e incluso las de tipo C). Otras en cambio se limitan a sólo uno de estos conjuntos.

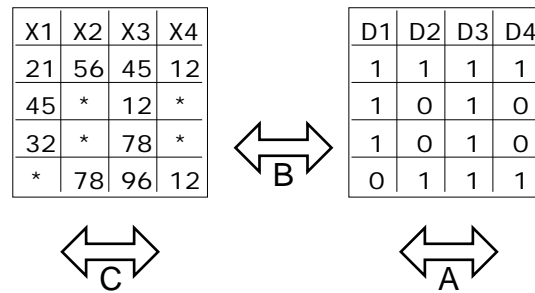


Figura 6.11: Cuestiones a explorar en datos con valores faltantes.

Mostraremos a continuación algunas de las técnicas que es posible utilizar siguiendo el orden de su complejidad.

### 6.6.1. Exploración univariada de datos faltantes

Este es el caso más simple. A este nivel nos estaremos preguntando por el número y proporción o porcentaje de los datos faltantes. También puede resultar interesante examinar los datos presentes con objeto de apreciar si los presentes son plausibles o si se ha producido un cierto "recorte" respecto a lo esperado.

MANET (<http://www1.math.uni-augsburg.de/Manet/>) es un programa destinado fundamentalmente al análisis gráfico de datos faltantes. Entre sus aportaciones, de las cuales veremos varias más a lo largo de los diferentes apartados de esta sección, se encuentra una modificación de los tradicionales histogramas, diagramas de barras y gráficos de cajas. Un ejemplo, utilizando los datos de Raymond (1983), es el siguiente. En el lado izquierdo se representa el histograma de la variable EDAD. Las barras grises corresponderían con un histograma usual al que se añade, en el lado izquierdo, una barra blanca para indicar el número de valores faltantes para esa variable. El otro gráfico correspondería a la variable PREV (número de cursos de idiomas cursados anteriormente) y se trata de un diagrama de barras, el cual también incluye una barra adicional para los valores faltantes. Esto ha sido mostrado en la sección anterior.

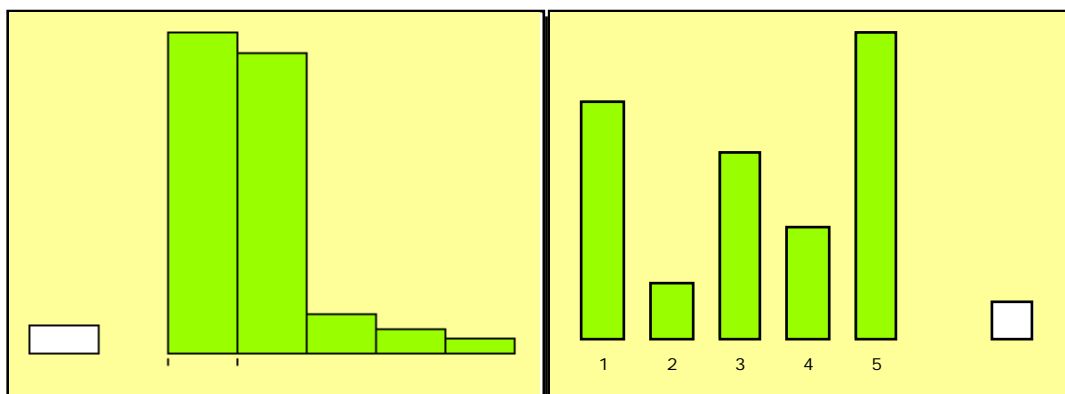


Figura 6.12: Gráficos estadísticos modificados para incluir datos faltantes.

Otros programas son capaces de mostrar gráficos incluyendo los valores faltantes aunque, a diferencia de MANET, en muchos casos no lo hacen automáticamente y es necesario un pequeño esfuerzo adicional para obtenerlos.

Por último, un resultado que es interesante también examinar es el referido a los datos descriptivos básicos (media, desviación típica, etc.) para los datos completos frente a los datos disponibles. De este modo, podemos tener una indicación del daño que se produciría si optáramos por la vía más sencilla de eliminar aquellos casos que tuvieran valores faltantes.

### 6.6.2. Exploración bivariada/multivariada de datos faltantes

Lo siguiente corresponde a una matriz de correlaciones <sup>4</sup> para las variables con valores faltantes de los datos de Raymond (1997) después de haberlos convertido en valores binarios en donde 1 indica faltante y 0 presente. Tenemos por tanto una representación que nos permite observar las relaciones de tipo A aunque no las de tipo B. Se han señalado en negrita aquellos valores que indican un alto grado de solapamiento. En general podemos ver que hay dos grupos de variables que coinciden entre sí. En primer lugar TALM, TAEV, TAEM e ING, y, en segundo, EDAD con PREV. Un inconveniente que presenta esta matriz de correlaciones es que los valores no informan acerca del número de casos implicados en el cálculo de esa correlación (aunque podría incluirse en las filas y columnas). Una alternativa sería incluir la proporción de solapamiento entre casos faltantes o simplemente la frecuencia.

<sup>4</sup> Estas correlaciones corresponden al caso de la Phi de Pearson al estar basadas en variables dicotómicas.

	EDAD	PREV	SEXO	TALM	TAEV	TAEM	ING	CALI	CALA	NING
EDAD	1.000									
PREV	<b>1.000</b>	1.000								
SEXO	-0.012	-0.012	1.000							
TALM	0.116	0.116	-0.027	1.000						
TAEV	0.049	0.049	-0.022	-0.079	1.000					
TAEM	0.049	0.049	-0.022	-0.079	<b>1.000</b>	1.000				
ING	0.041	0.041	-0.023	-0.091	<b>0.952</b>	<b>0.952</b>	1.000			
CALI	-0.012	-0.012	-0.004	-0.027	0.164	0.164	0.156	1.000		
CALA	0.049	0.049	-0.022	-0.079	<b>1.000</b>	<b>1.000</b>	<b>0.952</b>	0.164	1.000	
NING	0.018	0.018	0.135	0.232	0.076	0.076	0.117	-0.027	0.076	1.000

*Tabla 6.13. Matriz de correlaciones para variables ficticias*

Una representación similar en contenido a la matriz de correlaciones es la mostrada en la tabla 6.14 y está disponible en BMDP. En ella se muestran los diferentes patrones de valores faltantes de una manera condensada y un recuento del número de veces que aparece ese patrón. Las celdas en blanco indican valor presente y las que tienen un 1 valor faltante. En el recuadro se puede apreciar el solapamiento entre los valores de LAN y EDAD que producía un correlación de 1 como veíamos anteriormente.

MANET incluye gráficos dinámicos como método para explorar esas interrelaciones. Los gráficos dinámicos (Cleveland and McGill, 1988) parecen una de las últimas incorporaciones al análisis de datos, posibilitado fundamentalmente por el desarrollo de ordenadores con grandes capacidades gráficas y con una interacción ágil, que permita la exploración de la información. Otros programas que permiten el uso de gráficos dinámicos son DataDesk (<http://www.datadesk.com>) y MacSpin. El gráfico 6.15 es un ejemplo de este tipo de métodos y permite una exploración similar en objetivos a lo comentado hasta ahora.

LEPSETTTICCN					174
ADREAAAANA AI				1	18
NAEXLLEEGLLN				1	1
DVOEMVM IAG				1	2
				1 1 1	1
				1 1 1	20
				1 1 1	7
				1 1 1 1	1
				1	26
				1	1
				1 1 1 1	15
				1 1 1 1	2
				1 1 1 1	1
				1	1
				1 1	3
				1 1	1
				1 1	1 1 1
				1 1	2
				1 1	1
				1 1	3
				1 1	1
				1 1	1
				1 1 1 1	1 1
				1 1	1

Tabla 6.14. Representación de patrones de valores faltantes

Cada barra representa el total de datos. En ella es posible ver tres tonalidades (colores en la pantalla). La parte negra indica los valores presentes y la blanca, los faltantes. La parte gris es la parte *seleccionada*, es decir, que hemos marcado previamente con el ratón. En este ejemplo, la parte marcada ha sido los valores faltantes de TAE-V. Por ello, aunque esa parte debería estar en blanco, ha cambiado a gris. Debido a que las diferentes barras están interconectadas entre sí, aparecen partes grises en todas las variables. Cuando la parte gris está en el lado izquierdo significa que los valores faltantes en TAE-V no coinciden con los valores faltantes en la variable respectiva (por ejemplo, LAN o EALE). Cuando la parte gris está en el lado derecho, entonces cubre la parte blanca, simbolizando que existe solapamiento. A veces, naturalmente, la coincidencia será parcial y aparecerá un parte gris sobre la parte blanca y otra sobre la parte negra.

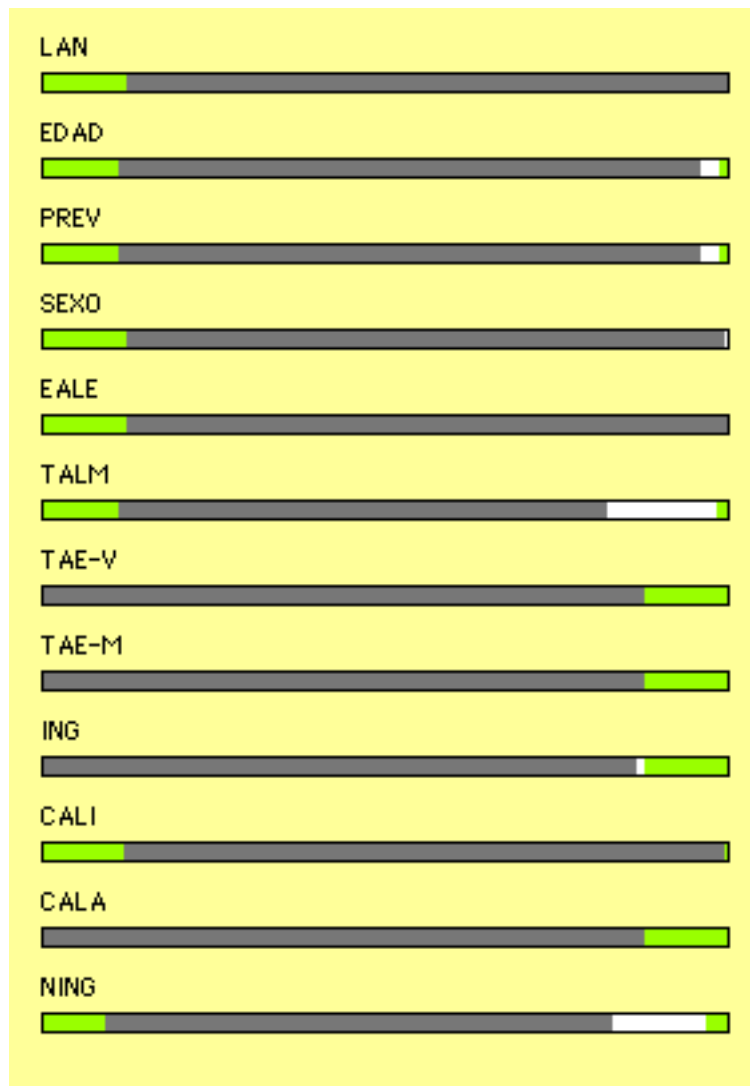


Figura 6.15: Gráfico de datos faltantes en MANET

En nuestro caso, podemos ver fácilmente que TAE-V coincide sobre todo con TAE-M, ING y CALA.

Los gráficos y técnicas presentados hasta ahora presentan el inconveniente de estar excesivamente centrados en el problema de los datos faltantes y olvidan la relación entre éstos y los valores en otras variables (lo que denominábamos relaciones tipo B). El siguiente gráfico presenta un avance en esa dirección. Este es un diagrama de dispersión que presenta dos añadidos. Unas puntos blancos sobre los ejes que representan la ausencia de un valor para una de las variables, y unas barras blancas que indican la proporción de valores faltantes en la variable situada en el eje x, en ambas variables simultáneamente, y, finalmente sólo en el eje y. En Manet, este gráfico está

interconectado con los anteriores por lo que es posible realizar una serie de comprobaciones adicionales en caso de ser necesario. En nuestro caso, los puntos en gris son los correspondientes a los valores faltantes en la variable NING. Es posible ver que todos ellos coinciden con los valores más bajos en ambas variables por lo que podemos inferir que, en caso de utilizar los datos completos de estas tres variables en un análisis la correlación entre TALM y TAE-V se vería disminuida, aunque no excesivamente (en concreto pasa de 0.417 a 0.403).

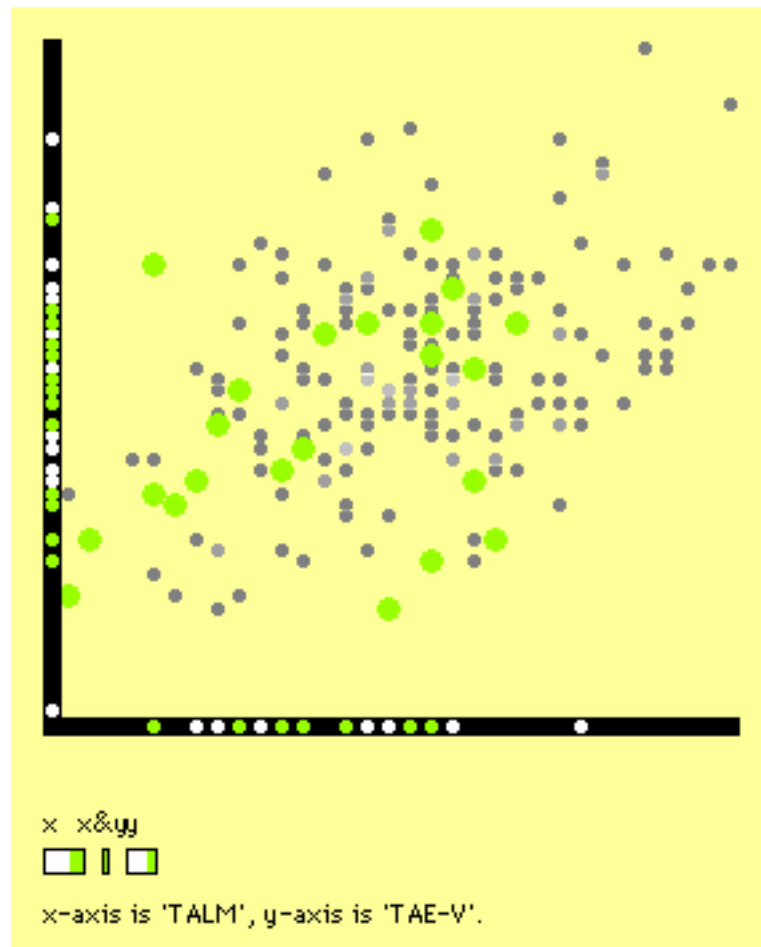


Figura 6.16: Diagramas de dispersión con datos faltantes señalados

L E P S E T T T I C C N A D R E A A A A N A A I N A E X L L E E G L L N D V O E M V M I A G		LAN				TAE-V	TAE-M	ING	
		1	2	3	4				
		174	41	50	83	0	511.328	575.063	55.005
	1	18	3	3	4	8	473.333	515	47.111
	1	1	1	0	0	0	590	680	.
	1 1	2	0	1	0	1	330	420	.
	1 1 1 1	20	8	4	8	0	.	.	.
	1 1 1 1 1	7	0	4	1	2	.	.	.
	1 1 1 1 1 1	1	0	1	0	0	.	.	.
	1	26	8	8	10	0	512.692	553.077	50.576
	1 1	15	0	3	4	8	485.333	532.667	53.266
	1 1 1 1 1 1	2	0	1	1	0	.	.	.
	1 1 1 1 1 1 1	1	0	1	0	0	.	.	.
	1	1	1	0	0	0	600	520	57
	1 1	3	1	0	1	1	413.333	560	50.666
	1 1	1	0	0	0	1	340	440	51
	1 1 1 1 1 1	2	2	0	0	0	.	.	.
	1 1 1	3	1	0	2	0	543.333	616.667	69.333
	1 1 1 1	1	0	0	1	0	400	530	49
	1 1 1 1 1 1 1 1	1	0	1	0	0	.	.	.

Tabla 6.17: Datos descriptivos por patrón de datos faltantes

Si siguiendo con la idea de mostrar las relaciones entre valores faltantes y las otras variables, el módulo de análisis faltantes de SPSS (White Paper, 1997) incluye el resultado de la tabla 6.17 (aproximadamente), que puede considerarse una ampliación del



presentado anteriormente. En ella puede verse como los estadísticos descriptivos de las variables se desglosan en función de los diferentes patrones de valores faltantes en un esfuerzo por describir simultáneamente variables discretas y variables continuas. Así, para la variable discreta LAN podemos ver el número de sujetos que se encuentra en cada una de sus categorías para cada patrón de valores faltantes, mientras que para las variables continuas TAE-V, TAE-M e ING vemos las medias.

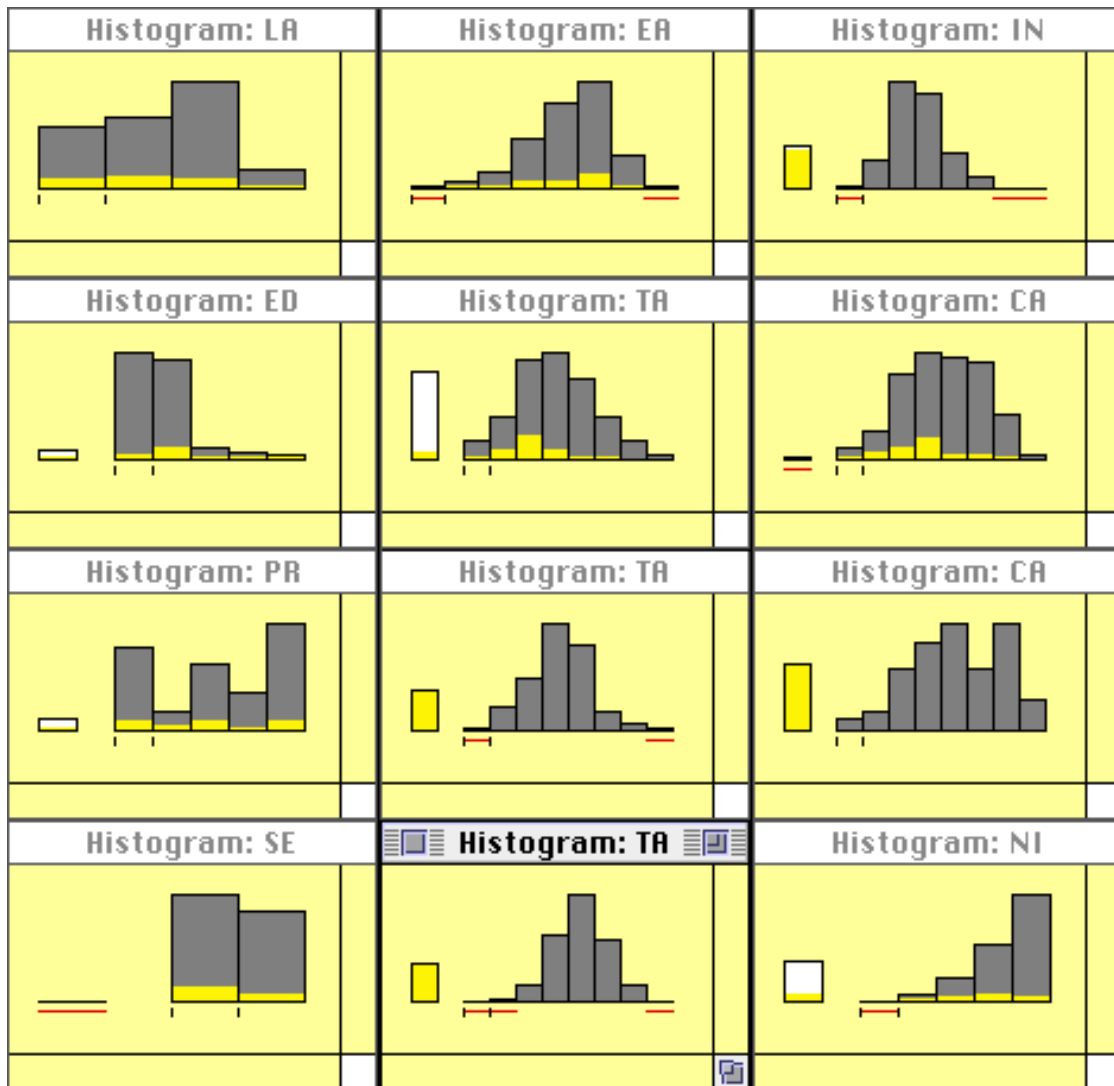


Figura 6.18: Histogramas conectados para analizar la interconexión entre valores faltantes

De este modo, tenemos que, por ejemplo, ninguno de los sujetos que tiene como lenguaje el ruso (LAN=4) posee los datos completos. Podemos ver que, de los 21 sujetos que estudian ruso, la mayoría parece coincidir que no tienen nota en inglés (variable NING). También podemos ver que la nota media para la variable de aptitud verbal (TAE-

V) es inferior en muchos casos para los diferentes patrones que la de los datos completos (primera fila).

MANET ofrece una alternativa gráfica a la tabla anterior. En el gráfico de la figura 6.18 podemos ver una matriz de histogramas de las diferentes variables<sup>5</sup>. Esta matriz es una representación gráfica por medio de la cual podemos obtener una impresión de la distribución de las diferentes variables. Seleccionando un grupo de datos obtenemos un nuevo histograma superpuesto al anterior en color gris. De este modo obtenemos una impresión gráfica de lo seleccionado. En este caso, hemos seleccionado los valores faltantes en la variable TAE-M con el resultado mostrado en la figura 6.18.

## **6.7. Asignación de valores faltantes**

Puesto que la eliminación de aquellos casos con valores faltantes presenta bastantes inconvenientes resulta natural buscar alternativas que eviten este extremo. Una solución en apariencia natural, aunque no exenta a su vez de peligros, es asignar un valor razonable a los huecos existentes en nuestros datos para, a continuación, realizar los análisis como si éstos fuera los valores reales.

A continuación revisaremos métodos para realizar esta asignación. Empezaremos en primer lugar por los denominados "rápidos" (Little and Rubin, 1987) para luego pasar a otros más complicados y, en general, más correctos. Antes de ello, queremos presentar un argumento que justifica las ventajas de llevar a cabo asignaciones.

En los datos acerca de enseñanza de un lenguaje extranjero veíamos que faltaban aproximadamente un 7% de los valores de la tabla, pero que su utilización con técnicas multivariadas suponía anular aproximadamente un 30% del total de los casos. Esto supone que los análisis utilizando aquellos casos que tengan todos los datos disponibles son un 70% correctos con respecto al total de los datos. Supongamos que asignamos al 7% de valores faltantes otros que son, digamos, un 80% correctos (por ejemplo, tomando el valor de unidades similares en el resto de los variables y que no son faltantes en la variable considerada). Los datos totales serían ahora  $100 - (7 \cdot 0.8) = 94.4\%$  correctos. Naturalmente, existen métodos de asignación de datos que pueden hacer más daño que

---

<sup>5</sup> Aunque lo apropiado para una variable discreta es un diagrama de barras (por ejemplo, para la variable LA), las diferencias gráficas son poco importantes en nuestro caso y no justifican el esfuerzo añadido.

beneficio pero cuando se eligen alternativas lo suficientemente válidas las ganancias pueden ser considerables.

Nos ceñiremos a los métodos basados en el supuesto que podemos ignorar los mecanismos que han producido la no-respuesta y que, por tanto, los datos son faltantes completamente al azar o al menos faltantes al azar. Es necesario recalcar que, debido a que no existe información acerca de estos supuestos en los datos, resulta tarea del investigador determinar su implausibilidad.

### 6.7.1 Métodos rápidos

En este apartado discutimos métodos que están disponibles fácilmente cuando se utiliza un paquete estadístico ordinario. Algunas de estas opciones son incluso ofrecidas automáticamente (p.e. asignación del valor medio de la variable en el modulo REGRESSION de SPSS) o son calculables con un conocimiento moderado de los lenguajes de programación incluidos en ellos.

- Asignación del valor medio de la variable: Este es probablemente el método más sencillo de asignación de valores faltantes y consiste en estimar los valores faltantes  $y_{ij}$  por  $\bar{y}_j^{NF}$ , la media de los valores no faltantes. Esto hace que la media de los valores observados y asignados sea  $\bar{y}_j^{NF}$ , que es la misma que la media de los análisis de casos disponibles. Este método presenta el inconveniente de que, puesto que los valores asignados están siempre en el centro de la distribución, se produce una infraestimación de la varianza verdadera (que corresponde a  $[n^{NF} - 1] / [n - 1]$ ). Lo mismo ocurre con las covarianzas, que son infraestimadas por un factor de  $[n^{NF_j NF_k} - 1] / [n - 1]$ . No obstante, las matrices de varianzas-covarianzas sí son positivas definidas. Correcciones a estas varianzas y covarianzas revierten a fórmulas similares a las discutidas en las secciones dedicadas al análisis de valores faltantes usando valores disponibles.

- Asignar el valor medio condicional sobre los valores de otras variables (método de Buck): Ciñéndonos al caso más habitual de  $y_k$  variables siguiendo la distribución normal-multivariada entonces nos encontramos con el caso de la regresión lineal de las variables observadas sobre las variables con casos faltantes (Buck, 1960). Este método presenta las siguientes consecuencias negativas:

a) Varianzas reducidas: Ciñéndonos al caso bivariado, si partimos de que la varianza de una variable con valores faltantes se puede descomponer en  $\sigma_j^2 = \sigma_{jk}^2 + \sigma_k^2$ , en donde  $\sigma_{jk}^2$  es la pendiente de la regresión, y, por tanto  $\sigma_{jk}^2$  correspondería a la varianza explicada por la variables sobre la que se ha condicionado, y,

$y_{j,k}$  sería la no explicada por esa variable, entonces vemos que la variabilidad de las puntuaciones faltantes, que fueron asignadas mediante este método, tendrían menor varianza total, ya que la segunda parte sería exactamente cero. Esto se produce porque naturalmente, las puntuaciones asignadas son puntuaciones predichas-ajustadas, sin ninguna variabilidad debida al error, y, por tanto, sobreajustadas con respecto a las correspondientes, no observadas, puntuaciones empíricas. En concreto, la infraestimación se produciría por la cantidad  $(n - n^j)(n - 1)^{-1} y_{j,k}$ . Es conveniente hacer notar que la infraestimación es pequeña cuando la predicción es buena ya que la proporción de varianza no explicada sería pequeña.

b) Covarianzas sesgadas: Puesto que los datos estimados caen en una línea recta respecto del predictor es lógico suponer que la covarianza se verá cambiada. Por ejemplo, (Little and Rubin, 1987) si tenemos dos variables con datos faltantes  $y_k$  e  $y_j$ , los cuales son estimadas utilizando otras variables, el sesgo en la covarianza entre esas variables será  $(n - 1)^{-1} c_{jk,i}$ , en donde  $c_{jk,i}$  es igual a la covarianza residual de estas variables tras haberlas regresado sobre los valores observados en las otras variables cuando ambas eran simultáneamente faltantes.

Las estimaciones del sesgo en las covarianzas y medias podrían ser utilizadas para corregir las estimaciones en una especie de paso iterativo, pero esto se aproxima a los métodos de máxima verosimilitud que discutiremos más adelante.

Estos métodos basados en regresión pueden extenderse mediante la utilización de variables ficticias a los casos en que las variables sin valores faltantes son categoriales. No obstante, cuando estas variables tienen valores faltantes y deben adoptar el papel de dependientes las estimaciones lineales pueden no caer en las categorías de 1 y 0 y por tanto otros métodos de estimación resultan más convenientes, tal y como la regresión logística (Aldrich, 1984).

A pesar de sus inconvenientes, este método supone una mejora sobre la utilización de medias incondicionales ya que la magnitud de los errores de estimación de la varianza son menores y puede ser puesto en relación con los métodos más correctos que describiremos a continuación. Un problema práctico que nos llevará a utilizar paquetes estadísticos diseñados específicamente para tratar con estos problemas es el de la creación de regresiones lineales para cada patrón de valores faltantes, tarea que puede resultar sumamente complicada con los programas comúnmente utilizados.

- Asignación estocástica por medio de regresión. Consiste en reemplazar los valores faltantes por regresión más un residual, el cual es extraído para incorporar falta de

exactitud en el valor predicho. En regresión, este residual debería tener media cero y varianza igual a la varianza residual de la regresión. En el caso binario se utilizaría regresión logística, que produciría la probabilidad de un valor de ser cero o uno, el cual sería extraído de la población con esa probabilidad. Una desventaja obvia de este método es que, puesto que la distribución del error no es necesariamente homogénea condicional sobre los predictores, el residual tendería a infraestimar la posible heterogeneidad de la varianza residual. Una opción a considerar para evitar este problema es extraer los valores a sustituir a partir de las unidades que sí han respondido, método que veremos a continuación.

### **6.7.2 Métodos basados en selección de otros candidatos.**

El tema común en estos métodos es, dado un caso de no respuesta, seleccionar un caso similar en las variables disponibles a partir de otros sujetos que sí han respondido (Little and Rubin, 1987) o bien buscar otros que sustituyan a los que no contestan. Empezaremos por este último caso en primer lugar porque pensamos que puede resultar lo más habitual en estudios prácticos.

- Sustitución de unidades que no responden por otros no seleccionados previamente en la muestra: Este método plantea el problema que las unidades que son sustituidas difieren de aquellas que aceptaron hacerlo desde un principio. Esto puede sesgar la muestra y, en cualquier caso, los análisis deberían incorporar que las unidades seleccionadas en la muestra son, en cierto modo, valores asignados. Del mismo modo, toda la información disponible acerca de aquellas unidades que se han negado a responder debería ser registrada y analizada como parte del estudio.

Los otros métodos son los siguientes:

- Extraer los valores a asignar a partir de otros casos que sí han respondido en el mismo estudio (Hot Deck) que sean similares al que no lo ha hecho con respecto a las variables disponibles.

- Extraer los valores de otros estudios o de una fuente externa (Cold Deck).

Métodos de este tipo son bastante habituales en la construcción de datos censales. En este caso, los creadores de la base de datos están obligados a proporcionar una única información completa que pueda ser sometida a una variedad de análisis por diferentes investigadores.

Rubin (1987) describe el procedimiento utilizado por la Agencia del censo de los Estados Unidos. Este consiste en:

1. Hacer todas las variables X categóricas.
2. Encontrar un donante que encaje exactamente con los valores categoriales de las variables disponibles respecto al caso no-respondente.
  - 2.1. Si aparecen varios donantes exactamente iguales al buscado se selecciona según el método en particular utilizado: a) el primero de los encontrados o b) uno extraído al azar de los encontrados.
  - 2.2. Si no aparecen donantes, algunas de las categorías de las variables disponibles se hacen más amplias (por ejemplo, si tenemos una categoría por cada estado podríamos agruparlas en función de si son del Sur o del Norte). Todos los casos faltantes recibirán finalmente información para completarlos. No obstante, se aplicarán diferentes reglas en función de como de fácil sea encontrar casos completos que se ajusten al perfil deseado.

En la práctica, a menos que se utilice un método para ampliar el espectro de posibles donantes tal y como el descrito para el censo resultará muy difícil encontrar uno que corresponda exactamente a lo buscado por nosotros. Una forma de manejar esto puede ser utilizando un método *Hot Deck* métrico.

• Métodos métricos *Hot Deck*: Estos métodos (Rubin, 1987) definen una medida de distancia entre los respondentes y los no respondentes tal y como por ejemplo:

$$d(i, i') = \max |x_{ij} - x_{i'j}|$$

En donde tenemos J variables escaladas apropiadamente (por ejemplo en puntuaciones típicas) para las que  $i'$  tiene un valor faltante en otra variable Y. A continuación podríamos elegir un valor para la variable faltante a partir de un grupo de candidatos que hubieran contestado en la variable Y y cuya distancia respecto del no respondente no superara un valor  $d$ . El número de candidatos puede ser controlado variando el tamaño de  $d$ . Otras definiciones de distancia son también posibles.

Otro método para la definición de distancias es el conocido por "emparejamiento de puntuaciones de proclividad a la falta de respuesta" (Rubin, 1987). Por ejemplo SOLAS 1.0 (<http://www.statsol.ie/>) sigue el siguiente algoritmo:

1) Se crea una variable temporal que será usada como variable dependiente en un modelo de regresión logística. Esta variable será 0 por cada caso en la variable que estemos asignando que sea faltante y 1 en otro caso. SOLAS permite seleccionar las variables covariantes a partir de las que se realizará la asignación.

2) Usando los coeficientes de regresión calculamos la proclividad de ser faltante. Esta puntuación es por tanto la probabilidad condicional de ser faltante, dado el vector de covariantes.

3) A continuación se dividen las puntuaciones de proclividad en quintiles. Dentro de cada quintil se hace un recuento del número de valores faltantes y presentes.

4) Para cada quintil se extrae una muestra aleatoria, con reemplazamiento de las respuestas observadas, igual en número a los datos observados. Esto crea la distribución predictiva posterior de la variable de interés.

5) Se toma una segunda muestra, con reemplazamiento, igual en tamaño al número de valores faltantes, y se usa esta muestra para asignar los valores faltantes.

### **6.7.3 Métodos combinados.**

Otra opción disponible consiste en combinar algunos de los métodos anteriores para llevar a cabo las asignaciones. De hecho, el método utilizado por SOLAS puede ser considerado de este tipo. Una idea similar es utilizar asignaciones basadas en regresión y entonces añadir un residual elegido aleatoriamente de los residuales respecto de los valores predichos al formar valores para la asignación. Este método (Graham, 1998) es el utilizado por EMCOV (<http://methcenter.psu.edu/EMCOV.html>) dentro del contexto de estimación mediante el algoritmo EM que introduciremos a continuación.

### **6.7.4 Métodos basados en máxima verosimilitud.**

Los métodos basados en máxima verosimilitud (Eliason, 1993) suponen una estrategia de estimación de parámetros en situaciones donde métodos más tradicionales resultan inadecuados. En el caso de la estimación de valores faltantes, podemos plantear el problema en los siguientes términos. Asumiendo que los datos son Faltantes al Azar, nuestro objetivo al estimar los datos faltantes sería maximizar<sup>6</sup> (Rubin, 1987):

---

<sup>6</sup> Por razones de cálculo normalmente el valor que se toma para maximizar es el logaritmo de la verosimilitud. No es el caso aquí.

$$L(\theta / Y_{pres}) = \int f(Y_{pres}, Y_{falt} / \theta) dY_{falt}$$

En donde  $\theta$  son los parámetros a maximizar, los cuales, en el caso de variables normales multivariadas serán normalmente las medias y las matrices de varianzas-covarianzas, y que no corresponden solamente a los datos observados sino a *todos* los datos.  $Y_{pres}$  es la parte observada de los datos y  $Y_{falt}$  es la parte no observada. Como es posible ver, esta verosimilitud depende no sólo de los valores presentes sino también de los faltantes teniendo en cuenta los parámetros de los datos. En general, no existe una solución explícita para los parámetros por lo que una solución implícita es normalmente necesaria. Esa solución proviene de herramientas especiales tal y como el algoritmo EM (o el de aumento de datos considerado posteriormente). EM se apoya en la interdependencia entre datos faltantes y parámetros ya que el hecho que los datos faltantes contengan información relevante para estimar los parámetros y estos a su vez contengan información importante para estimar los valores faltantes son la base del esquema de estimación que sigue este algoritmo, el cual es descrito a continuación.

Rubin (1991) señala que la idea detrás de EM es muy vieja e intuitiva y puede ser descrita del siguiente modo:

1. Dado un problema que es difícil de resolver <sup>7</sup>, formularlo de modo que si los datos faltantes fueran observados entonces la solución sería directa. En concreto, formular el problema de tal modo que una buena estimación (p.e. la estimación de máxima verosimilitud) de los parámetros  $\theta$ ,  $\hat{\theta}$  serían fáciles de hallar si los valores faltantes  $Y_{falt}$  fueran observados tal y como los valores observados  $Y_{obs}$ .
2. A continuación, asignar unos valores a  $Y_{falt}$  y solucionar el problema (es decir, encontrar  $\hat{\theta}$ ).
3. Usar este  $\hat{\theta}$  para encontrar los mejores valores de  $Y_{falt}$ , y entonces repetir el punto 2 para encontrar un nuevo valor de  $\hat{\theta}$ .
4. Iteracionar hasta que los valores de  $\hat{\theta}$  converjan.

---

<sup>7</sup> Existen casos en que la forma de los datos permite la estimación de los datos faltantes sin recurrir a métodos iterativos. Para una descripción v. Schafer (1997) o Little y Rubin (1987). No serán descritos aquí por razones de espacio.



Little y Rubin (1987) mencionan otros algoritmos más conocidos para la estimación mediante máxima verosimilitud que podrían ser considerados para resolver este problema. Estos son *Newton-Raphson* y el algoritmo de puntuación (*scoring*). Ambos métodos ofrecen el inconveniente de necesitar el cálculo de la matriz de derivadas segundas de la verosimilitud (la cual permite determinar si el valor hallado mediante la primera derivada es un máximo o un mínimo -Eliason, 1993-). Este cálculo resulta bastante complicado cuando los patrones de datos incompletos son complejos y los métodos para su obtención necesitan una programación muy cuidadosa. El algoritmo EM no necesita en cambio obtener las segundas derivadas puesto que puede ser demostrado (Rubin, 1991) que cada paso del algoritmo mejora la verosimilitud de los parámetros hasta llegar a la convergencia.

Cada iteración de EM consiste de un paso E de Esperado (*Expectation*) y M (*Maximization*) de Maximización. El paso E encuentra los valores esperados condicionales de los valores faltantes dados los valores observados y los parámetros actuales, y entonces sustituye los datos faltantes por esos valores esperados. El paso M encuentra los valores de los parámetros dados los valores hallados en el paso E en cada iteración.

El algoritmo EM ha sido aplicado a datos de muchos tipos con datos faltantes: Datos con distribución normal bivariada o multivariada, tablas de contingencia múltiple y datos con variables normales y no-normales mezcladas. Veremos a continuación dos de los casos más comunes. El primero de ellos más sencillo (aunque no trivial), con datos bivariados normales, y el segundo con datos multivariados normales en general.

### *EM para datos bivariados*

Expondremos como ejemplo la aplicación de EM a un caso de datos bivariados con valores faltantes en ambas variables (Little and Rubin, 1987). El patrón a observar es el mostrado en la figura 6.19.

Tenemos un primer grupo de unidades con datos faltantes en la segunda variable y no en la primera, un grupo con valores presentes en ambas variables y un grupo final con datos faltantes en la primera variable y no en la segunda. Estamos interesados en calcular la estimación máximo-verosímil de la media  $\mu$  y la matriz de varianzas covarianzas de  $Y_1$  e  $Y_2$ . Esta tarea se puede simplificar si nos centramos en el cálculo de las siguientes estadísticas suficientes:

$$SUM_1 = \sum_{i=1}^n y_{i1},$$

$$sum_2 = \sum_1^n y_{i2},$$

$$sumc_1 = \sum_1^n y_{i1}^2,$$

$$sumc_2 = \sum_1^n y_{i2}^2,$$

$$sumprod_{12} = \sum_1^n y_{i1}y_{i2}$$

Estos corresponden a la suma de las variables, la suma de los cuadrados y la suma del producto. Las medias muestrales, las varianzas y las covarianzas son función de estos estadísticos.

$Y_1$	$Y_2$	
1	0	1=Observado 0=Faltante
.	.	
.	.	
.	.	
1	0	
1	1	
.	.	
.	.	
.	.	
1	1	
0	1	
.	.	
.	.	
.	.	
0	1	

Figura 6.19: Patrón de datos faltantes para datos bivariados.

Para encontrar los valores esperados de estos estadísticos, dados los valores observados, necesitaremos calcular algunos valores mientras que otros nos vienen dados directamente. Así, para el grupo de unidades con  $y_{i1}$  e  $y_{i2}$  observados los valores esperados coinciden con los observados. Para el grupo de  $y_{i1}$  observado pero  $y_{i2}$  faltante, los valores esperados de  $y_{i1}$ ,  $y_{i1}^2$  pueden obtenerse de los valores observados, mientras que los valores esperados de  $y_{i2}$ ,  $y_{i2}^2$  e  $y_{i1}y_{i2}$  deben ser hallados a partir de la regresión de  $y_{i2}$  sobre  $y_{i1}$ .

$$E(y_{i2} | y_{i1}, \mu) = \mu_{201} + \mu_{211}y_{i1}$$

$$E(y_{i2}^2 | y_{i1}, \mu) = (\mu_{201} + \mu_{211}y_{i1})^2 + \mu_{221}$$

$$E(y_{i2}y_{i1}|y_{i1}, \mu) = (\sigma_{201} + \sigma_{211}y_{i1})y_{i1}$$

En donde  $\sigma_{201}$ ,  $\sigma_{211}$  y  $\sigma_{221}$  son funciones de la matriz de varianzas-covarianzas. Para las unidades con  $y_{i2}$  observado e  $y_{i1}$  faltante se usaría la regresión de  $y_{i1}$  sobre  $y_{i2}$  para calcular los estadísticos suficientes. Habiendo encontrado los valores esperados de las observaciones, pasaríamos a obtener los valores esperados de los estadísticos suficientes anteriormente indicados.

El paso M calcularía los estimadores basados en momentos habituales a partir de los estadísticos suficientes para los datos asignados. Estos serían:

$$\begin{aligned} \hat{\mu}_1 &= \text{sum}_1/n \\ \hat{\mu}_2 &= \text{sum}_2/n \\ \hat{\sigma}_1^2 &= \text{sum}c_1/n - \hat{\mu}_1^2 \\ \hat{\sigma}_2^2 &= \text{sum}c_2/n - \hat{\mu}_2^2 \\ \hat{\sigma}_{12} &= \text{sum}p/n - \hat{\mu}_1\hat{\mu}_2 \end{aligned}$$

El algoritmo EM para este problema consistiría en ejecutar estos pasos iterativamente (Little and Rubin, 1987).

#### *EM para datos normales multivariados incompletos*

Muchos análisis estadísticos, tal y como el análisis de regresión múltiple, análisis de componentes principales, análisis discriminante y correlación canónica se basan en un resumen inicial de los datos que consiste en obtener la media muestral y la matriz de covarianzas de las variables de la matriz de datos. Por ello, realizar esta estimación en el caso de muestras con datos incompletos resulta especialmente importante. A continuación se mostrarán métodos para llevar a cabo esta estimación, asumiendo que los datos son una muestra normal incompleta multivariada y que los valores faltantes lo son al azar. Aunque la restricción de normalidad multivariada resulte un tanto estricta, esta puede ser relajada y por tanto partir de supuestos más débiles en ocasiones (Little and Rubin, 1987). Schaffer (1997) por ejemplo señala las siguientes situaciones en las que un modelo normal puede resultar útil a pesar que se producen desviaciones de éste: 1) Cuando se pueden realizar transformaciones de los datos que hagan el supuesto de normalidad más aceptable, 2) Cuando se tienen variables claramente no-normales (p.e. discretas) pero han sido observadas completamente siempre y cuando sea plausible modelar las variables incompletas como condicionalmente normales dada una función de

las completas y los parámetros de interés inferencial pertenezcan sólo a esta distribución condicional. 3) Finalmente, incluso cuando las variables observadas incompletamente son claramente no-normales todavía puede resultar plausible usar el modelo normal como un método conveniente para crear asignaciones múltiples, un tema que será tratado más adelante. Las asignaciones múltiples pueden ser robustas a desviaciones del modelo de asignación si las cantidades de información no son demasiado grandes, puesto que el modelo de asignación no es aplicado al conjunto de los datos, sino sólo a la parte faltante. Por ejemplo, puede ser razonable realizar imputaciones de una variable ordinal (que consista en un pequeño número de categorías ordenadas), a condición que la cantidad de datos faltantes no sea excesiva y la distribución marginal no esté demasiado lejos de ser unimodal y simétrica. Schafer (1997) señala que este método les ha permitido incluso asignar variables binarias cuando cualquier otro método hubiera resultado poco práctico.

Seguiremos a partir de ahora la descripción del método según Little y Rubin (1987). Supongamos que tenemos  $K$  variables  $(Y_1, Y_2, \dots, Y_k)$  que tienen una distribución normal con media  $\mu = (\mu_1, \dots, \mu_k)$  y matriz de covarianzas  $\Sigma = (\sigma_{jk})$ . Tenemos  $Y = (Y_{PRES}, Y_{FALT})$  en donde  $Y$  es igual a una muestra aleatoria de tamaño  $n$  sobre  $(Y_1, \dots, Y_k)$ ,  $Y_{PRES}$  son los valores presentes u observados y  $Y_{FALT}$  son los valores faltantes. Para derivar el algoritmo partimos de los estadísticos suficientes:

$$S = \sum_{i=1}^n y_{ij}, j = 1, \dots, K \quad \text{y} \quad \sum_{i=1}^n y_{ij}y_{ik}, j, k = 1, \dots, K$$

Encontrándonos en la iteración  $t$ , con  $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$  indicando los parámetros actuales, el paso E del algoritmo consistirá en:

$$E \sum_{i=1}^n y_{ij} | Y_{PRES}, \theta^{(t)} = \sum_{i=1}^n y_{ij}^{(t)}, j = 1, \dots, K$$

$$E \sum_{i=1}^n y_{ij}y_{ik} | Y_{PRES}, \theta^{(t)} = \sum_{i=1}^n (y_{ij}^t y_{ik}^t + c_{jki}^{(t)}), j, k = 1, \dots, K$$

En donde,

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{si } y_{ij} \text{ está presente} \\ E(y_{ij} | y_{PRES, i}, \theta^{(t)}) & \text{si } y_{ij} \text{ está faltante} \end{cases}$$

Y,

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{si } y_{ij} \text{ o } y_{ik} \text{ están presentes} \\ \text{Cov}(y_{ij}, y_{ik} | y_{PRES,i}^{(t)}) & \text{si } y_{ij} \text{ y } y_{ik} \text{ están faltantes} \end{cases}$$

Los valores faltantes  $y_{ij}$  son así reemplazados por la media condicional de  $y_{ij}$  dado el conjunto de valores  $y_{PRES,i}$  presentes para esa observación. Estas medias condicionales, y las covarianzas condicionales no nulas son encontradas fácilmente de los parámetros estimados actuales utilizando el procedimiento *SWEEP de la matriz de covarianzas aumentada* (descrito más adelante) que permite convertir las variables  $y_{PRES,i}$  en predictores en la ecuación de regresión y las variables restantes en variables predichas.

El paso M del algoritmo EM es directo. Las nuevas estimaciones  $^{(t+1)}$  de los parámetros son estimados de los estadísticos suficientes de los datos completos. Esto es:

$$\begin{aligned} \mu_j^{(t+1)} &= n^{-1} \sum_{i=1}^n y_{ij}^{(t)}, \quad j = 1, \dots, K; \\ c_{jk}^{(t+1)} &= n^{-1} E \sum_{i=1}^n y_{ij} y_{ik} | Y_{PRES} - \mu_j^{(t+1)} \mu_k^{(t+1)} \\ &= n^{-1} \sum_{i=1}^n \left[ (y_{ij} - \mu_j^{(t+1)}) (y_{ik} - \mu_k^{(t+1)}) + c_{jki}^{(t)} \right], \quad j, k = 1, \dots, K \end{aligned}$$

Faltaría por sugerir los valores iniciales de los parámetros. Existen cuatro posibilidades directas: 1) Usar la solución de casos completos; 2) usar una de las soluciones de casos disponibles discutidas en secciones anteriores; 3) formar la media muestral y la matriz de covarianzas por medio de uno de los métodos de asignación de datos faltantes considerados en esta sección anteriormente; 4) utilizar la media y las varianzas de los datos completos y situar las correlaciones a cero. En general todas las soluciones son aceptables, aunque algunas de ellas pueden llevar a problemas de estimación, así que un programa de ordenador que permitiera la estimación debería permitir diferentes opciones.

Según Little y Rubin (1987) el algoritmo aquí presentado se debe a Orchard y Woodbury (1972). El algoritmo de puntuación para este problema había sido presentado anteriormente por Trawinsky y Bargmann (1964) pero presenta problemas de cálculo que limitan su uso.

*El operador SWEEP (Barrer)*

El algoritmo EM anteriormente descrito puede parecer enormemente trabajoso incluso para un ordenador si se tiene en cuenta que cada variable con valores faltantes necesita aparecer como variable predicha en cada iteración de una ecuación de regresión que incluye los datos completos como predictores. Sin embargo, su cálculo se facilita enormemente cuando se utiliza el operador de matrices SWEEP, el cual proporciona los resultados básicos a partir de los que es posible obtener las ecuaciones de regresión correspondientes. Este operador fue según Little y Rubin (1987) propuesto originalmente por Beaton (1964).

El operador SWEEP es definido para matrices simétricas como sigue. Una matriz  $p \times p$  simétrica  $G$  es *barrida* con respecto a la fila y la columna  $k$  si es reemplazada por otra matriz simétrica  $p \times p$  denominada  $H$  con elementos definidos del siguiente modo:

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk}, \quad k \neq j, \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk}, \quad k \neq j, k \neq l \end{aligned}$$

La notación  $SWP[k]G = H$  indica que se ha *barrido* sobre la fila y columna  $k$ .

Este operador está muy relacionado con la regresión lineal. Supongamos que tenemos la siguiente matriz  $G$ :

$$G = \begin{matrix} 1 & \bar{y}_1 & \dots & \bar{y}_j & \dots & \bar{y}_p \\ \bar{y}_1 & n^{-1} & y_1^2 & & \dots & n^{-1} & y_p y_1 \\ \vdots & \vdots & \ddots & & & \vdots & \\ \bar{y}_k & & & n^{-1} & y_j y_k & \ddots & \\ \bar{y}_{p1} & n^{-1} & y_1 y_p & & \dots & n^{-1} & y_p^2 \end{matrix}$$

Haciendo  $SWP[0,1]G$  sobre esta matriz obtenemos lo siguiente:

$$\begin{array}{ccccccc}
 & -\left(1 + \bar{y}_1^2/s_{11}\right) & \bar{y}_1/s_{11} & \bar{y}_2 - (s_{12}/s_{11})\bar{y}_1 & \cdots & \bar{y}_p - (s_{1p}/s_{11})\bar{y}_1 & \\
 & & -1/s_{11} & s_{12}/s_{11} & \cdots & s_{1p}/s_{11} & \\
 SWP[0,1]G = & & & s_{22} - s_{12}^2/s_{11} & \cdots & s_{2p} - s_{1p}s_{12}/s_{11} & = \begin{array}{cc} -A & B \\ B^T & C \end{array} \\
 & & & & & \mathbf{M} & \\
 & & \bar{y}_p - (s_{1p}/s_{11})\bar{y}_1 & & & s_{pp} - s_{1p}^2/s_{11} & 
 \end{array}$$

En donde  $A$  es  $2 \times 2$ ,  $B$  es  $2 \times (p - 1)$  y  $C$  es  $(p - 1) \times (p - 1)$ . Esta matriz produce resultados para la regresión de  $Y_2, \dots, Y_p$  sobre  $Y_1$ . En concreto, la columna  $j$  de  $B$  da la intercepta y la pendiente para la regresión de  $Y_{j+1}$  sobre  $Y_1$  para  $j=1, \dots, p - 1$ . La matriz  $C$  da la matriz de covarianzas residuales de  $Y_2, \dots, Y_p$  dado  $Y_1$ . Finalmente, los elementos de  $A$ , cuando son multiplicados por la varianza o covarianza residual apropiada en  $C$  y divididos por  $n$ , producen varianzas y covarianzas de los coeficientes de regresión estimados en  $B$ .

Haciendo *SWEEP* por los primeros  $q$  elementos y la constante tenemos unos resultados similares a los mostrados en la matriz superior pero para una regresión multivariada en la que las primeras  $q$  variables toman el papel de predictores y las restantes el de predichas.

### Consideraciones finales

Los procedimientos anteriores permiten la obtención de los parámetros para los datos. Estos parámetros pueden ser utilizados para realizar los cálculos correspondientes a muchos análisis multivariados tal y como componentes principales, regresión, etc. Por ello, cuando la tarea del investigador se limite al ajuste de modelos puede ser suficiente obtener esos parámetros.

Sin embargo, en ocasiones resulta interesante llevar a cabo una asignación de datos dentro de la tabla original. Las razones para ello pueden ser el interés por utilizar programas estadísticos que sólo aceptan datos originales (y no parámetros) como comienzo de los análisis, o cuando el objetivo de la asignación es proporcionar matrices de datos a otros investigadores (tal y como en los conjuntos de datos de uso público como censos o datos de evaluación nacional, etc.).

En ese caso, tras haber calculado los parámetros y, a partir de ellos, los valores esperados condicionales correspondientes a los valores faltantes, una asignación de datos

correcta debería restaurar el error original en las variables. Esta restauración se puede realizar de dos maneras: 1) Reasignando variabilidad debida al error y 2) Muestreando la ecuación de regresión.

1) Restauración de la variabilidad debida al error.

Existen dos posibilidades igualmente aceptables:

- Muestreo de los residuales para los datos que no son faltantes. Este método es el utilizado por el programa EMCOV (Afifi and Clark, 1984).
- Añadir una variable normal aleatoria al valor asignado tal y como hace el programa NORM (Schafer, 1997).

Según Graham (1998), no hay una evidencia clara que haga preferir un método al otro (aunque por otras razones él recomienda utilizar el programa NORM antes que el desarrollado por él mismo).

2) Muestreo de la ecuación de regresión.

Existen también dos posibilidades.

- EMCOV utiliza un procedimiento de bootstrapping. Básicamente, aplican el método de bootstrap sobre los datos, para luego seguir con el algoritmo EM para producir una matriz de covarianzas y entonces utiliza esa matriz de covarianza para producir los valores asignados.
- NORM utiliza un procedimiento conocido como de "aumento de datos" (*data augmentation*). En este método los pasos determinísticos E y M son sustituidos por pasos estocásticos denominados I (*Imputation-Asignación*) y P (*Posterior-Distribución condicional posterior*). Básicamente, esto significa que en lugar de estimar un valor esperado para los valores y las matrices de varianzas covarianzas, se lleva a cabo una simulación que extrae los valores correspondientes de las distribuciones condicionales posteriores. Es decir, se simula la distribución de los valores esperados y se realiza una o varias extracciones para cada caso, en el paso I, o se construyen una o varias matrices de varianzas-covarianzas, en el paso P (Schafer, 1997). La extracción de varias matrices y conjuntos de datos convierten este método en una aplicación del método de asignación múltiple que describiremos a continuación.



### 6.7.5 Asignación múltiple

Los métodos descritos hasta ahora (salvo el procedimiento de *aumentación de datos* que ya anticipa esta sección) son denominados por Rubin (1987) como de asignación sencilla. Esto significa que cada valor faltante es sustituido por un único valor, restaurando al conjunto de datos su integridad. Esto presenta las siguientes ventajas:

- Permite aplicar métodos standard de datos completos. Puesto que muchos investigadores poseen y utilizan paquetes y métodos estadísticos diseñados para datos completos esta ventaja resulta especialmente interesante puesto que no tienen que recurrir a herramientas especializadas como las que necesitarían en caso de utilizar otros métodos para tratar con los datos faltantes.

- La asignación de datos puede incorporar conocimiento que sólo tiene el que recoge los datos pero que no estará disponible públicamente. Cuando se trata de bases de datos de uso público resulta habitual que la agencia u organismos encargado de publicar esos datos posea información que, debido a limitaciones legales, no puede ser distribuida. Sin embargo, a partir de ella, resulta posible utilizar esta información para llenar los datos desconocidos y de ese modo proporcionar bases de datos más completas.

La asignación múltiple es una técnica que reemplaza cada valor faltante o deficiente con uno o más valores aceptables que representan una distribución de posibilidades. Los inconvenientes de la asignación simple con respecto a esta estrategia serían las siguientes:

- Asignar un valor único por cada valor faltante y luego pasar a hacer análisis estadísticos significa tratar ese valor como si fuera conocido, por lo que, sin ajustes especiales, no es posible apreciar la inseguridad asociada al desconocimiento de esos valores.

- En ocasiones existen varios modelos de asignación de datos (en este texto no se ha tratado el tema de asignación de datos bajo modelos de no respuesta y siempre se ha hablado tal y como si ésta fuera *ignorable*) que podrían resultar plausibles. La asignación simple no tiene manera de tratar con esa inseguridad, mientras que la múltiple lo puede hacer de una manera relativamente natural.

En definitiva, la asignación múltiple trata la inseguridad asociada con la asignación de datos realizando repeticiones del mismo proceso para luego combinar los diferentes resultados de tal modo que se pueda obtener una apreciación de la consistencia de éstos. Métodos para combinar estas estimaciones, así como los errores típicos asociados serán

descritos más adelante. Antes comentaremos algunas de las desventajas de la asignación múltiple. Estos son:

- Realizar asignaciones múltiples necesita más trabajo que hacerlas simples. Este problema no es muy grave si se tiene en cuenta que en general un número modesto de asignaciones múltiples (entre 0 a 10 pero generalmente sólo 3) es considerado suficiente. Rubin (1991) muestra que la eficiencia relativa de la estimación usando un número infinito de asignaciones múltiples v. únicamente tres con un 30% de información faltante es sólo un 5% por ciento menor.

En general, los inconvenientes de este método provienen del aumento en el esfuerzo de análisis derivado de tener que realizar tres veces los mismos cálculos y almacenar archivos de datos mayores. Sin embargo, en los tiempos actuales esto no es excesivamente problemático en líneas generales debido a la disponibilidad de ordenadores capaces de gestionar problemas de gran tamaño con relativa facilidad.

- Combinación de las diferentes asignaciones (Naus, 1982, Rubin, 1987, Rubin, 1991, Schafer, 1997). Después de haber obtenido diferentes estimaciones para los datos faltantes y diferentes modelos un paso fundamental es realizar una combinación de estas estimaciones para apreciar como la variabilidad asociada a éstas resulta lo suficientemente importante como para poner en duda los parámetros obtenidos o, por el contrario, resulta poco importante y podemos por tanto confiar en las asignaciones realizadas.

Para estimar la estimación combinada de un parámetro con un único componente ( $r=1$ ) estimado un número  $m$  de veces podemos calcular la media:

$$\bar{I}_m = \frac{\sum_{l=1}^m \hat{I}_l}{m}$$

La variabilidad asociada con esta estimación tiene dos componentes: En primer lugar, el promedio de la varianza intra-asignaciones corresponde a la combinación del error asociado al parámetro para cada uno de los conjuntos de datos llenados:

$$\bar{I}_m = \frac{\sum_{l=1}^m \hat{I}_l}{m}$$

En segundo lugar está la variabilidad entre asignaciones:

$$E_m = \frac{\left( \hat{\sigma}_m^2 - \bar{\sigma}_m^2 \right)^2}{m - 1}$$

Sumando estos dos componentes se obtiene una estimación de la variabilidad total asociada con  $\bar{\sigma}_m$

$$T_m = \bar{I}_m + \frac{m + 1}{m} E_m$$

En donde  $m + 1/m$  es un ajuste para  $m$  finito. La distribución de referencia para estimaciones de intervalo y pruebas de significación es, siguiendo la distribución  $t$ :

$$\left( \bar{\sigma}_m \right) T_m^{-1/2} \sim t_v$$

Valores cercanos a cero indican que no hay una variabilidad muy grande entre los parámetros derivados de las diferentes asignaciones. Valores altos que superen los valores de referencia de la distribución  $t$  significarán que la variabilidad asociada a cada una de las asignaciones produce resultados sustancialmente diferentes .

En donde los grados de libertad son:

$$v = (m - 1) \left[ 1 + \frac{1}{m + 1} \frac{\bar{I}_m}{E_m} \right]^2$$

Este procedimiento puede ser extendido para parámetros con varios componentes ( $r > 1$ ), y también existen procedimientos para combinar directamente valores de probabilidad para pruebas de significación (Rubin, 1987).

## 6.8. Consideraciones finales

Este texto no es completo. La literatura sobre datos faltantes, a diferencia de la referida a otros temas en este texto es muy considerable. Además, existe una gran actividad en cuanto a proponer nuevos algoritmos y técnicas con este propósito. Por ejemplo, Rubin (1991) discute varios algoritmos de estimación de datos faltantes que aquí no han sido ni siquiera mencionados. También, en los últimos tiempos se han considerado las redes neuronales aplicadas a este caso (Gharamani y Jordan, 1994). Por último, muchas técnicas estadísticas son capaces de tratar con aspectos de datos faltantes mediante métodos propios que quizás merecerían ser discutidos. Dado el propósito de este texto, sin embargo, lo tratado hasta aquí parece suficiente.

Por otro lado, parece que estas técnicas se encuentran ya en un momento de madurez adecuado para convertirse en herramienta de uso habitual. De este modo, varios paquetes estadísticos de propósito general (p.e. SPSS, BMDP) están incorporando asignación de datos faltantes usando métodos de máxima verosimilitud como parte de su oferta. Esto es un avance importante, pues hasta ahora estos programas a menudo se limitaban a realizar un borrado *listwise* de modo automático, sin informar al usuario de lo que estaba ocurriendo, arrojándose en el prestigio de su marca (si SPSS lo hace estará bien hecho...?).

Sin embargo, hay una barrera que costará probablemente un poco de saltar. Es la referida a la sensación que muchos revisores pueden tener que estas técnicas consisten básicamente de "inventar" datos, es decir, de engañar de alguna forma. Para cambiar esta perspectiva, opinamos que las técnicas de visualización de datos pueden ayudar a hacer entender el funcionamiento de las técnicas y, por tanto, a distinguir entre usos legítimos o ilegítimos de aquellas. Este camino ya ha sido emprendido en nuestro grupo de investigación y está empezando a producir sus primeros frutos en tiempos recientes.





# 7. ***Transformación y generación de datos***

A menudo, los datos tal y como son obtenidos directamente no tienen la forma adecuada para ser analizados. Por tanto, es necesario *transformarlos* previamente. Tal y como señala Davidson (1996), puesto que los paquetes estadísticos son capaces de realizar esas transformaciones rápida y eficientemente, resulta conveniente planear el proceso de los datos para aprovechar al máximo esta característica. Esto garantiza además una mayor precisión en los resultados.

No obstante, las transformaciones pueden ser a su vez una fuente de errores: los valores faltantes pueden no ser considerados correctamente, las fórmulas utilizadas pueden ser escritas incorrectamente o algunos casos especiales no ser considerados. Así, los métodos de detección de errores discutidos en secciones anteriores también deberían ser utilizados tras realizar una transformación para evitar que los fallos pasen inadvertidos.

Independientemente de los posibles errores, es indudable que las transformaciones son una herramienta de gran utilidad. A modo de resumen podemos indicar las siguientes funciones:

- Puesto que a menudo los datos están separados en diversos archivos y/o fuentes es necesario hacer manipulaciones que proporcionen unidad a la información. También, en ocasiones, el problema es el contrario y lo que se desea es dividir la información en partes coherentes y que puedan ser manejadas con comodidad. Finalmente, también tenemos el caso en que los datos han sido recogidos al mismo nivel que se quieren hacer los análisis, por lo que es necesario colapsar o agregar la información, obteniendo recuentos u otros resúmenes de la información tal y como medias, etc. para las unidades consideradas.

- Las variables, tal y como son recogidas, no siempre se encuentran en un estado adecuado. Por ejemplo, pueden ser excesivamente complejas, difíciles de interpretar de una manera directa o con limitaciones desde un punto de vista estadístico en cuanto a su análisis. Precisamente, puesto que realizar transformaciones puede modificar estos aspectos en las variables, el encargado de realizarlas deberían tener en cuenta consideraciones sustantivas que garanticen la corrección del resultado. Asimismo, puede ser conveniente combinar variables para formar otras compuestas (por ejemplo, cuando se hacen sumas de variables). Por último, puede ser conveniente para ciertas tareas producir variables nuevas que seguirían una distribución aleatoria bajo un modelo de probabilidad aleatoria (uniforme, Poisson, etc.).

- Por último, en ocasiones queremos hacer transformaciones que consideran los valores de modo individual y que suponen decisiones para cada de uno de ellos.

Las transformaciones aquí comentadas pueden ser llevada a cabo mediante diversos programas. El encargado de procesar datos probablemente manejará uno u otro en función de la complejidad del problema, la flexibilidad necesaria y otras consideraciones. En este caso hemos asumido implícitamente que la transformación se está realizando con un paquete estadístico. Sin embargo, un programa orientado a la gestión de bases de datos o una hoja de cálculo pueden ser capaces de cumplir muchas de las funciones anteriormente comentadas de modo similar. No obstante, es responsabilidad del encargado del proceso de datos evaluar hasta qué punto el programa utilizado realiza correctamente las tareas deseadas. Por ejemplo, algunas bases de datos rellenan los valores vacíos con ceros (Davidson, 1986), mientras que ciertas hojas de cálculo sumarán un valor faltante a uno correcto sin producir ninguna nota de alerta (p. e. Excel 5. 0). Ambos comportamientos son inadecuados y pueden generar alteraciones graves en los datos originales.

En el resto de la sección, se presentan diversos tipos de transformaciones de datos que suelen utilizarse frecuentemente en la práctica y que, normalmente, vienen integradas



en las funciones ofrecidas por los programas al uso en la actualidad. Las transformaciones consideradas están clasificadas en referidas a ficheros completos, o a variables dentro de ficheros. Dentro de este último apartado dedicaremos una sección importante al concepto de reexpresión, el cual es marco que encuadra las justificaciones para cierto tipo de transformaciones. Finalmente, señalaremos algunas soluciones que se han propuesto para el problema del registros de las transformaciones realizadas para evitar confusiones entre los diversos ficheros que, inevitablemente, se generan durante el procesamiento de los datos.

## 7.1. Manipulación de ficheros

Generalmente, las operaciones sobre ficheros van agrupadas por pares, siendo una la inversa de la otra. Consideraremos las siguientes operaciones de manipulación de archivos.

- **Creación de un fichero a partir de otro.** Consiste en generar un fichero de datos a partir de uno ya existente, bien sea por selección, bien por eliminación, de algunas de las variables o casos del fichero de datos de partida. Normalmente, podremos asignar un nuevo nombre al fichero así generado, con lo que dispondremos tanto del fichero original como del generado a partir de éste. Normalmente, se incluyen opciones para, durante el proceso, eliminar algunas variables y conservar otras, de tal modo que el nuevo fichero puede limitarse a una parte especial del conjunto.

- **Eliminación de ficheros.** Normalmente, ningún paquete estadístico proporciona comandos para eliminar un fichero completamente desde el propio programa. Esto normalmente debe realizarse desde el propio sistema operativo, utilizando funciones por tanto externas al programa.

- **Unión de ficheros.** Consiste en unir los datos de dos o más ficheros, creando un nuevo fichero que contendrá los datos de todos los ficheros unidos. Esta unión o pegado se puede realizar de alguna de las dos siguientes formas:

1. Pegado horizontal. Podemos distinguir dos situaciones. En la primera, los dos archivos tienen los mismos casos mientras que en la segunda se unen ficheros con niveles diferentes.

*Cuando cada caso en un fichero se corresponde con un caso en el otro fichero.* Supone reunir en un mismo fichero, la información correspondiente a las variables de dos o más ficheros. Tendrá sentido este tipo de unión en el caso en que la

información de las variables corresponda a los mismos casos y éstos estén ordenados del mismo modo en todas las variables. En el ejemplo de la figura 7.1 puede verse una representación que haría referencia a dos ficheros. El fichero izquierdo tiene cuatro variables y el derecho cinco. El objetivo del pegado vertical sería unir ambos ficheros para crear uno nuevo con las nueve variables. Naturalmente, esta unión sólo es legítima si cada caso (cada fila) corresponde a un mismo sujeto. Para que esto sea así ambos ficheros deben estar ordenados del mismo modo. Aunque en este caso no sea estrictamente necesario, una precaución adicional muy interesante es introducir para cada fichero una *variable identificadora* para cada sujeto en cada fichero. Esta variable nos permitirá comprobar que efectivamente la unión se ha realizado correctamente comparando sus valores en el nuevo fichero.

SEXO	Y-PMA	R-PMA	N-PMA	N-EPQ	E-EPQ	P-EPQ	S-EPQ	CA-EPQ
1	24	17	10	10	17	1	16	18
2	24	17	10	10	17	1	16	18
1	26	15	12	15	23	10	10	26
1	26	15	12	15	23	10	10	26
2	35	12	17	8	21	1	15	17
2	31	12	6	10	21	1	13	18
2	21	14	16	12	23	2	10	22
1	31	16	18	4	21	2	10	14
2	31	16	18	4	21	2	10	14
2	27	18	21	15	12	3	10	18
1	5	15	17	8	18	2	13	16
2	18	14	15	13	22	7	16	24
1	33	12	11	14	15	2	21	21
2	25	18	27	12	23	5	17	23
1	26	18	11	8	11	4	9	13

Figura 7.1: Representación de pegado horizontal

Quando cada caso en un fichero se corresponde con varios casos en otro fichero. En ocasiones la información de cada caso en un fichero corresponde con la información de varios casos en otro fichero. Esta situación ocurre muy habitualmente por ejemplo en situaciones educativas en las que se tiene información acerca de estudiantes que están dentro de escuelas. Parte de la información puede corresponder a la escuela (tal y como el presupuesto que ésta tiene, su tamaño, etc.) y parte al estudiante (su rendimiento académico, su nivel socioeconómico). Esta estructura es denominada multinivel o jerárquica (Bryk, 1992), desde un punto de vista estadístico y como relacional cuando se trata de una base de datos. En la parte izquierda de la figura 7.2 se muestra una tabla en que cada caso corresponde con un alumno y otra en la parte derecha en el que cada caso corresponde a una escuela. En esta segunda tabla se indica el número de alumnos por clase, mientras que en el

primero se indica a qué escuela pertenece cada alumno así como variables referidas al rendimiento académico. El resultado de la transformación aquí comentada sería asociar a cada alumno el número de estudiantes que hay por clase en su colegio.

Fichero 1				Fichero 2	
<b>Escuelas</b>	<b>VALENCIANO</b>	<b>LENGUA</b>	<b>IDIOMA</b>	<b>Escuelas Al. clase</b>	
Escuela 5	3	2.5	4	<b>Escuela 1</b>	30
Escuela 2	2	2	3	<b>Escuela 2</b>	18
Escuela 4	1	1.5	1.5	<b>Escuela 3</b>	20
Escuela 2	2	2.5	2.5	<b>Escuela 4</b>	58
Escuela 3	3.5	4	4	<b>Escuela 5</b>	19
Escuela 3	4	4	4		
Escuela 4	1.5	1.5	0.5		
Escuela 2	2	4	1.5		
Escuela 5	3	1.5	2		
Escuela 2	2	1.5	2		
Escuela 4	1	1	0.5		
Escuela 4	1	3	2		
Escuela 5	2.5	1	4		
Escuela 4	1.5	1	0.5		
Escuela 2	2	2	3		
Escuela 4	1	1.5	2		

Figura 7.2: Pegado horizontal con variables a niveles diferentes

Es conveniente comentar que esta transformación es poco eficiente desde el punto de vista del uso de los recursos de un ordenador. Produce una nueva variable que consume más espacio que utilizar una *relación* para una ambas tablas. Davidson (1996) opina que es un truco inadecuado pero práctico. En realidad, si se parte del hecho que la mayoría de los paquetes estadísticos están diseñados para trabajar con estructuras con forma de tabla esta solución resulta prácticamente inevitable. Una discusión más completa de esta estructura de datos se encuentra en el capítulo uno, en la parte referida a estructuras de datos y computadores.

2. Pegado vertical. En esta situación se parte de dos archivos que corresponden a las mismas variables pero con filas referidas a casos diferentes. Esta puede ser la situación por ejemplo cuando la tarea de crear el fichero de datos ha sido dividida en dos o más bloques. También, cuando la información proviene de diferentes unidades de información (colegios, empresas, ciudades, etc.) y es necesario combinarla.

En la figura 7.3 se representa esta situación. Originalmente hay dos archivos con las mismas variables y el objetivo es lograr que la información en uno de ellos esté dispuesta "debajo" de la información en el otro.

SEXO	Y-PMA	R-PMA	N-PMA	SEXO	Y-PMA	R-PMA	N-PMA
1	24	17	10	1	24	10	24
2	24	17	10	2	40	11	18
1	26	15	12	1	25	19	23
1	26	15	12	2	16	12	10
2	35	12	17	1	23	12	14
2	31	12	6	1	30	22	21
2	21	14	16	2	22	19	13
1	31	16	18	2	23	16	18
2	31	16	18	2	18	18	28
2	27	18	21	1	20	19	13
1	5	15	17	2	30	17	22
2	18	14	15	2	31	20	16
1	33	12	11	1	31	20	16
2	25	18	27	2	23	10	18
1	26	18	11	2	21	12	13

Figura 7.3: Pegado vertical

• **Dividir ficheros de datos.** A menudo estamos interesados en dividir nuestros archivos para realizar análisis sobre parte de nuestro archivo. Esto puede ser conveniente por ejemplo para disminuir las variables en nuestro fichero de datos y así hacer más sencillo el ejecutar nuevos comandos o guardar la información. También, en ocasiones querremos seleccionar algunos casos atendiendo a criterios específicos o simplemente en función de una variable de grupo. Veremos estos dos casos por separado.

1. División Horizontal. Cuando separamos unas variables del archivo general y creamos un nuevo archivo con sólo parte de ellas estamos haciendo la operación inversa a la de la figura 7.2 (pegado horizontal con mismo número de casos). Esta operación puede ser entendida como una función de *exportación* en la que los datos exportados tienen el mismo formato informático que los datos originales, ya que, generalmente, durante el proceso de división, el programa elegido nos dará la opción de seleccionar las columnas o variables que queremos guardar.

La operación de guardar un archivo con menor número de casos que el original (operación opuesta al pegado horizontal con archivos de diferente número de casos) es equivalente a llevar a cabo una operación de agregación de datos, una operación que será descrita más adelante.

2. División vertical. En esta situación los casos de un archivo son separados dando lugar a dos o más archivos. Esta operación puede realizarse de dos maneras principalmente:

- División en función de una variable de grupo: En los datos de Raymond (1997) utilizados en el apartado de valores faltantes acerca de la enseñanza de lenguas extranjeras, los sujetos podían estudiar uno de cuatro lenguajes tal y como se indicaba en la variable LAN (Lenguaje Extranjero estudiado: 1=Francés; 2=Español; 3=Alemán; 4=Ruso). El encargado del proceso de datos podría dividir este fichero en cuatro partes, que tendrían la información para cada uno de los idiomas. Esta situación correspondería a una división por grupo.

- División en función de una selección de los casos: La separación anterior resulta sencilla porque tenemos una variable que es capaz de indicarnos cuales son los nuevos archivos a crear. En ocasiones, el criterio de separación no será tan visible y el investigador deseará utilizar otros más complicados.

La función *SI Condición<valor de variable> ENTONCES Seleccionar* permite esta separación. Por ejemplo, si en los datos de Raymond tuviéramos interés en separar la información correspondiente a los sujetos con notas altas (mayores que 100 por ejemplo) en la variable ING (puntuaciones altas en inglés en el momento de ingreso en la universidad) podríamos escribir:

*SI ING>100 ENTONCES Seleccionar*

Una discusión completa de las pruebas lógicas aplicables a variables y de la función *SI ENTONCES* queda pendiente para otro lugar.

Un tipo de selección especial es el que podríamos denominar *muestreo*. Esto corresponde con la situación en la que se selecciona, generalmente mediante un proceso aleatorio, un grupo de observaciones del total. Una aplicación de esta operación es, por ejemplo, seleccionar la mitad de los datos para llevar a cabo análisis exploratorios seguidos a continuación de otros confirmatorios con la otra mitad. Para ello, el propio ordenador genera valores aleatorios o pseudo-aleatorios que designan los casos elegidos.

- **Transposición.** Consiste en trasponer la matriz de datos, convirtiendo las columnas en filas y viceversa. De este modo, los valores de la primera columna pasarán a ser los correspondientes de la primera fila de la matriz de datos, y así sucesivamente.

Muchos paquetes estadísticos toman a las variables o columnas del fichero de datos como unidad de análisis. Consecuentemente, la transposición será un tipo de transformación del fichero de datos útil cuando se quiera analizar la información correspondiente a los casos o filas, ya sea porque se han introducido los datos al revés o porque los análisis estadísticos se quieran centrar sobre los casos tras un análisis de las

variables. La figura 7.4 muestra una representación gráfica de esta situación. En la parte izquierda se encuentra la tabla con el formato habitual de variables por sujetos y en la parte derecha el resultado de su transposición.

SUJ	VAL.	LEN.	IDIOMA	Hª	MATEM	C.N.	Variables	1	2	3	4	5	6
1	3	2,5	4	2,5	2	3	VAL.	3	2	1	2	3,5	4
2	2	2	3	2,5	3	2,5	LEN.	2,5	2	1,5	2,5	4	4
3	1	1,5	1,5	2	1	1	IDIOMA	4	3	1,5	2,5	4	4
4	2	2,5	2,5	2,5	1	1,5	Hª	2,5	2,5	2	2,5	3	2,5
5	3,5	4	4	3	1	3,5	MATEM	2	3	1	1	1	4
6	4	4	4	2,5	4	4	C.N.	3	2,5	1	1,5	3,5	4

Figura 7.4: Representación gráfica de una transposición

• **Ordenación.** Ordenar un fichero atendiendo a los valores de una variable puede ser útil por diversas razones: facilita la revisión de los datos, permite un examen superficial de éstos y, a veces, es una precondition para transformaciones más sofisticadas. La mayoría de los programas incorporan opciones para ordenar un fichero en función de una variable o varias de forma anidada. En la figura 7.5 es posible ver en la parte superior derecha el resultado de ordenar el fichero situado en la parte superior izquierda en función de la variable SEXO. Generalmente es posible realizar ordenaciones en dos sentidos, de menor a mayor o de mayor a menor.

En la parte inferior izquierda de la misma figura se puede ver el fichero ordenado simultáneamente por el Sexo, las notas en Valenciano y en Lengua. Esto se denomina ordenación *anidada* y se puede distinguir entre variables principales y variables secundarias con respecto a la ordenación. Las variables principales serán aquellas que permanecen ordenadas para todo el fichero. Las variables secundarias sólo estarían ordenadas dentro de las repeticiones de las variables principales. En nuestro caso, la variable SEXO es la principal y Valenciano y Lengua serán las secundarias.

En la parte inferior derecha es posible ver un error que suele ocurrir al realizar ordenaciones. La ordenación es la misma que en la parte izquierda pero se ha fallado en *arrastrar* las dos últimas columnas (por lo que permanecen iguales a la situación original, parte superior izquierda). Este error tienen consecuencias muy graves en caso de pasar inadvertido porque significa confundir la información para los casos.

A veces es necesario realizar ordenaciones atendiendo a más variables de las que el programa utilizado es capaz de manejar. En este caso se puede realizar la ordenación del siguiente modo. En primer lugar, se ordena por la variable menos importante arrastrando todas las demás. Repetir el paso anterior con el resto de las variables en orden decreciente de importancia hasta hacerlo en último lugar con la que consideramos principal. El resultado final es una ordenación anidada.

SEXO	VALENCIANO	LINGUA	IDIOMA	H2	SEXO	VALENCIANO	LINGUA	IDIOMA	H2
1	3	2.5	4	2.5	1	3	2.5	4	2.5
2	2	2	5	2.5	1	1	1.5	1.5	2
1	1	1.5	1.5	2	1	2	2.5	2.5	2.5
1	2	2.5	2.5	2.5	1	2	4	1.5	2.5
2	3.5	4	4	3	1	1	1	0.5	1.5
2	4	4	4	2.5	1	2.5	1	4	2
2	1.5	1.5	0.5	1.5	1	2	2	3	2
1	2	4	1.5	2.5	1	1	1.5	2	1.5
2	3	1.5	2	3.5	1	3	2	3.5	2
2	2	1.5	2	2	2	2	2	3	2.5
1	1	1	0.5	1.5	2	3.5	4	4	3
2	1	3	2	2	2	4	4	4	2.5
1	2.5	1	4	2	2	1.5	1.5	0.5	1.5
2	1.5	1	0.5	2	2	3	1.5	2	3.5
1	2	2	3	2	2	2	1.5	2	2
1	1	1.5	2	1.5	2	1	3	2	2
2	4	3.5	4	2	2	1.5	1	0.5	2
1	3	2	3.5	2	2	4	3.5	4	2
2	1	1	0.5	2	2	1	1	0.5	2

SEXO	VALENCIANO	LINGUA	IDIOMA	H2	SEXO	VALENCIANO	LINGUA	IDIOMA	H2
1	1	1	0.5	1.5	1	1	1	4	2.5
1	1	1.5	1.5	2	1	1	1.5	3	2.5
1	1	1.5	2	1.5	1	1	1.5	1.5	2
1	2	2	2	2	1	2	2	2.5	2.5
1	2	2.5	2.5	2.5	1	2	2.5	4	3
1	2	4	1.5	2.5	1	2	4	4	3.5
1	2.5	1	4	2	1	2.5	1	0.5	1.5
1	3	2	3.5	2	1	3	2	1.5	2.5
1	3	2.5	4	2.5	1	3	2.5	2	3.5
2	1	1	0.5	2	2	1	1	2	2
2	1	3	2	2	2	1	3	0.5	1.5
2	1.5	1	0.5	2	2	1.5	1	2	2
2	1.5	1.5	0.5	1.5	2	1.5	1.5	4	2
2	2	1.5	2	2	2	2	1.5	0.5	2
2	2	2	3	2.5	2	2	2	3	2
2	3	1.5	2	3.5	2	3	1.5	2	1.5
2	3.5	4	4	3	2	3.5	4	4	2
2	4	3.5	4	2	2	4	3.5	3.5	2
2	4	4	4	2.5	2	4	4	0.5	2

Figura 7.5: Ejemplos de ordenación de variables

• **Agregación.** La agregación es una operación que implica agrupar casos juntos, para posteriormente asignar uno o varios valores por cada grupo formado. Esta operación es en definitiva la inversa al pegado horizontal cuando un fichero tiene menos casos que el otro discutida anteriormente. Usando el ejemplo de las escuelas, hablaríamos de agregación cuando creemos un archivo en que cada caso sería una escuela junto a otras variables que corresponderían los estudiantes dentro de esa escuela. Por ejemplo, la media de rendimiento de los estudiantes, el número de estudiantes por aula, la proporción de hombres y mujeres, etc.

Algunas de las funciones agregadas puede ser: suma, media, variabilidad, máximo, mínimo, número de casos, etc.

La agregación es una transformación que implica consecuencias de gran importancia, por lo que no debería ser emprendida sin tener en cuenta la problemática implicada. Básicamente, la agregación supone un cambio de nivel al que nos estamos refiriendo al hacer nuestras preguntas, lo cual puede resultar en que las respuestas encontradas al nuevo nivel no puedan ser extrapoladas al nivel previo. Este peligro es bien conocido desde hace bastante tiempo. Thorndike (1939) escribió un artículo titulado.

No obstante, este problema sigue apareciendo en investigaciones actuales (Walker, 1993). En pocas palabras, la agregación fuerza a los datos a variar en sólo una dirección, dejando de lado la variación que puede haber dentro de las unidades que han sido colapsadas en un valor único. Por ejemplo, la relación entre rendimiento académico e inteligencia para cada estudiante dentro de una serie de escuelas presentará unos valores diferentes que la relación entre el rendimiento medio y la inteligencia media de los estudiantes para cada escuela. Generalmente esta agregación implica que los coeficientes de varianza explicada se hinchen de tal modo que los resultados aparezcan más llamativos (aunque la reducción en el tamaño de la muestra puede actuar en modo opuesto con respecto a la probabilidad).

En ocasiones el investigador se planteará preguntas legítimas que implican la relación entre variables que se encuentran a niveles diferentes entre sí. Por ejemplo, el presupuesto de una escuela puede ser puesto en relación con el rendimiento académico después de haber considerado el efecto de la inteligencia individual de un alumno. Esta estructura es conocida como multinivel o jerárquica y los métodos estadísticos para su análisis han experimentado un desarrollo muy considerable en los últimos tiempos (Bryk, 1992, Goldstein y McDonald, 1988, Kreft, 1995).

- Disgregación. La disgregación sería la transformación opuesta a la de agregación. Hay que tener en cuenta que, puesto que la información a nivel individual no está disponible, varias unidades recibirán el mismo valor. Por ejemplo, cada estudiante en un aula recibirá como número de alumnos en clase el mismo valor. Este tipo de transformación, del mismo modo que la agregación, debería pues ser meditada cuidadosamente.

## **7.2. Manipulación de variables**

Dentro de este apartado veremos en primer lugar las transformaciones que afectan a una variable, la generación de variables (aleatorias y de otros tipos) y las que afectan a dos o más variables. A continuación pasaremos a una sección que considera cuestiones acerca de la justificación de las transformaciones en contextos estadísticos, y que tienen un carácter más avanzado (y una extensión mayor).

### **7.2.1. Transformaciones que afectan a una variable**

- Transformación del tipo de variable: Algunos programas permiten la transformación del tipo de formato de una variable a otro diferente, siempre que se trate de un cambio coherente. Por ejemplo, una variable de tipo numérico puede ser cambiada a una variable



de tipo texto o viceversa. También, una variable de tipo fecha puede ser convertida en un número (y viceversa) ya que la mayoría de los programas codifican internamente las fechas con números. Así, el número 0 en una hoja de cálculo bien conocida corresponde a 1/1/1904.

- **Recodificación.** La recodificación supone cambiar los valores de una variable por otros. Podemos distinguir dos casos: 1) Cuando tengamos una variable categorial y 2) cuando sea continua.

*Cuando tengamos una variable categorial* la recodificación indicará los valores viejos y los nuevos por los que hay que sustituirlos. Por ejemplo, si hemos codificado la variable Sexo con valores 1 y 0 podemos querer sustituirlos por 1 y 2.. Escrito en pseudocódigo tendríamos:

```
Recodificar Sexo (0=1) (1=2)
```

Tener en cuenta que la recodificación debería hacerse secuencialmente, empezando por un lado del fichero y terminando por el otro. En caso contrario los valores de 1 nuevos podrían ser confundidos con los valores 1 viejos y la variable acabaría teniendo sólo valores 2. Esta recodificación es muy común cuando se ha realizado un estudio en el que había diversas categorías de respuesta y algunas de ellas han sido utilizadas muy poco por lo que resulta conveniente combinarlas en una nueva que las agrupe. La recodificación es de hecho un paso fundamental en análisis estadísticos llevados a cabo dentro de ciertos contextos (p.e. observación seguida de análisis secuenciales).

*Cuando recodificamos una variable continua*, esta es dividida en intervalos. Un ejemplo es cambiar una variable indicando Edad por tres grupos. En pseudocódigo haríamos:

```
Recodificar Edad (0 hasta 18)=1 (19 hasta 30)=2 (31 hasta 40)=3...
```

Hay que hacer notar que se asume que no hay valores intermedios tal y como 18.5. En ese caso, la recodificación debería modificarse para incluir los valores intermedios.

Esta transformación supone perder información con respecto a la variable por lo que será necesario justificarla adecuadamente en función de la situación.

- **Función de transformación.** Supone aplicar una misma función de carácter matemático a todos los datos de una variable. Los programas que suelen utilizarse para realizar transformaciones en los datos -hojas de cálculo y paquetes estadísticos, fundamentalmente- suelen incorporar un repertorio de funciones que varía según el

programa, pero que suele incluir funciones tales como: raíz cuadrada, exponenciación, logaritmo decimal, logaritmo natural, redondeo de decimales, truncamiento de decimales, valor absoluto, resto del cociente, seno coseno, tangente, etc.

Aparte de las funciones de carácter básico enumeradas, también suelen presentarse otras funciones más complejas, siendo algunas de ellas de uso frecuente en el campo de la psicología:

- Jerarquización: supone asignar un rango a cada uno de los valores de una variable cuantitativa, transformándola de este modo en ordinal.

- Puntuaciones diferenciales: esta transformación representa restar a cada uno de los valores de la variable, la media de esos valores.

- Puntuaciones típicas: igual que la de las puntuaciones diferenciales, pero además dividiendo por la desviación típica del total de los valores.

- Puntuaciones normalizadas: supone transformar los valores de modo que la distribución de los mismos sea normal.

- Puntuación centil: esta transformación nos proporcionará de cada dato, un tanto por ciento que representa el porcentaje de valores en esa variable cuya magnitud es inferior a la de ese dato. Este tipo de puntuación se utiliza mucho para expresar las medidas de los sujetos en una determinada variable, ya que informa de un modo sencillo de la posición de cada uno de los sujetos respecto al resto del grupo.

• Inversión de variables: En la inversión los valores altos de la variable pasar a ser los bajos y viceversa. Util para poner las puntuaciones en una escala en la misma dirección. Una fórmula podría ser:

$$y_i = (\max(X) - x_i + \min(X))$$

En donde  $y_i$  es la variable transformada.  $\max(x)$  es el máximo de la variable a transformar y  $\min(x)$  el mínimo.

### **7.2.2. Generación de variables**

Con este tipo de transformación se hace referencia a la creación de variables, no a partir de los datos de otras variables, sino por otros procedimientos. Estos son fundamentalmente dos: la generación de variables cuyos datos constituyen series numéricas de diversa complejidad y la generación de variables cuyos datos son aleatorios.

- **Series.** Se trata de variables que toman un valor para el primer caso como punto de partida, obteniéndose todos los demás aplicando un regla que parte de él. La serie más sencilla empieza en 1 y sigue añadiendo 1 (1, 2, 3...). De este modo, el segundo valor se obtiene a partir del primero, el tercero a partir del segundo y así, sucesivamente hasta obtener una variable que está constituida por una sucesión o serie de valores ligados entre sí.

Este tipo de variables nos puede resultar útil para otros fines o simplemente para enumerar las filas de la tabla de datos y poder determinar la posición de datos concretos con más precisión.

También a menudo es posible crear series más complejas en las que cada valor se repite un número fijo de veces. Estas variables son útiles para la generación de codificaciones para ciertos análisis (p.e. análisis de varianza).

- **Aleatorias.** La generación de los valores que configuran estas variables toma punto de partida un valor, al que se suele llamar semilla y que normalmente es determinado por el usuario o, en su defecto, por el propio ordenador. A partir de este valor se obtiene el primer número aleatorio, y a partir de éste el segundo, etc. Véase Perea y Pitarque (1990) para una revisión de algoritmos de generación de números pseudoaleatorios uniformes y siguiendo diversas distribuciones de probabilidad.

### 7.2.3. Transformaciones sobre 2 o más variables

Las transformaciones sobre dos o más variables permiten combinar información para cada caso. De nuevo, aquí sólo será tratado de una manera muy superficial, aunque existen cuestiones en este tipo de transformaciones relativamente complejas y que deberían ser tratadas con atención.

- **Combinación de variables.** Este tipo de transformación hace referencia a la combinación de los valores de dos o más variables, para crear una nueva variable cuyos valores son obtenidos a partir de la aplicación de algún tipo de operador aritmético entre los valores de las variables a combinar. Los operadores aritméticos habitualmente utilizados son la suma, la resta, la multiplicación y la división.

Si, por ejemplo, tenemos un fichero de datos que contiene las puntuaciones de un grupo de sujetos en un test, de modo que cada columna o variable representa las puntuaciones de los sujetos en cada ítem, podríamos crear una columna que resultará de la suma de todas ellas, con lo que para cada caso se obtendría la suma de las puntuaciones

obtenidas en cada uno de los ítems. Esta nueva variable representaría la puntuación total de cada uno de los sujetos en el test.

Hay que advertir que estas combinaciones no siempre son admisibles y que es posible cometer errores conceptuales graves si no se presta debida atención a ciertas cuestiones. Por ejemplo, realizar una suma de ítems para cuestionarios que claramente no son unidimensionales sería en principio incongruente, siendo necesario hacer estas combinaciones dentro de escalas. También, podría cuestionarse hasta que punto un ítem debe de puntuar exclusivamente sólo en una escala o podría ser repartido de forma ponderada entre varias (requiriéndose otros métodos tal y como el análisis factorial para hacer la combinación). Un último problema son los valores faltantes, los cuales deberían ser tenidos en cuenta y también tratados correctamente.

- **Conteo.** Este tipo de transformación sirve para crear una variable numérica que contiene para cada caso, el número de veces que aparece un determinado valor en un conjunto de variables. Para llevar a cabo esta transformación habrá que especificar tanto el valor cuya ocurrencia se quiere contar, como las variables a través de las que se quiera realizar el conteo.

- **Fusión de variables.** Este tipo de transformación será útil cuando tengamos dos o más variables y queramos reducirlas a una sola que ofrezca la misma información que cada una de las variables por separado. Este tipo de transformación será factible con variables nominales que tengan un número limitado y determinado de categorías.

Sean por caso, dos variables nominales, una con 3 categorías y la otra con 2. La información de estas dos variables podría ser expresada por una sola variable con 6 categorías, una por cada una de las combinaciones posibles entre las categorías de las dos variables.

### **7.3. Concepto de reexpresión**

El término reexpresión puede ser considerado como equivalente al de transformación, de tal modo que algunos autores lo utilizan de modo intercambiable (Emerson and Stoto, 1983) Sin embargo, el término reexpresión añade el matiz de que la transformación altera las propiedades de la escala en que los datos originales fueron medidos revelando cualidades que estaban ocultas previamente. Se busca generalmente hacer los análisis estadísticos más sencillos o mas claros al modificar la forma de los datos originales y sustituirla por una más adecuada.

No obstante, transformar la escala en que los datos originales están expresados implica inconvenientes. Así, los nuevos datos pueden ser más complicados de entender y de manejar, y en muchas ocasiones puede ser aparentemente poco razonable aceptar las relaciones establecidas que sólo se muestran sobre los datos transformados.

Una justificación teórica puede ayudarnos a aceptar con más facilidad las posibilidades que nos ofrecen estas técnicas. Esta es reconocer que las ciencias sociales no han desarrollado todavía un cuerpo de teoría lo suficientemente perfeccionado como para que las apropiadas escalas y relaciones entre dimensiones puedan ser determinadas, por lo que explorar otras escalas puede ser una buena fuente de avances en esa dirección. En la medida en que buenas teorías vayan apareciendo relaciones que actualmente nos parecen extrañas pueden mostrarse como naturales.

Otra justificación es la basada en argumentos estadísticos. Muchas técnicas estadísticas requieren de datos que se comporten según ciertas normas: las variables deberían ser simétricas y/o normales, la distribución de los errores debería ser constante, falta de relación entre variabilidad y categorías, etc. No obstante, a menudo los datos recogidos no se ajustan a estos requisitos: los salarios de los empleados en una empresa suelen presentar asimetría positiva (muchos cobran sueldos bajos, unos pocos cobran sueldos altos), la insatisfacción con productos de consumo altamente vendidos presenta asimetría negativa (muchos individuos lo encuentran bueno, sólo unos pocos lo rechazan completamente), etc.

Las transformaciones pueden convertir unos datos con esas características en otros que se comporten de la manera requerida por ciertas técnicas estadísticas.

Por último, señalar que, en realidad, existen muchas variables que manejamos en nuestra vida cotidiana que son transformadas de una manera rutinaria (Emerson, 1991) sin que despierte extrañeza (por ejemplo litros de gasolina por 100 kms, grados Celsius).

Una justificación más completa del interés de las transformaciones y su utilización deberá esperar a la exposición de sus casos más avanzados y tendrá que esperar al final de esta sección.

A continuación mostraremos una clasificación de las transformaciones apropiadas para variables numéricas, así como algunas de sus propiedades y efectos sobre los datos originales. Finalmente, se mostrará un ejemplo de como una transformación es capaz de revelar una relación que permanecía oculta a primera vista y, por último, se discutirán los argumentos a favor y en contra de su utilización.

### 7.3.1. Tipos de transformaciones.

Lo siguiente está basado en Emerson (1991). Distinguiremos entre transformaciones lineales y no lineales, y dentro de estas últimas en monótonicas y no monótonicas, así como en crecientes y decrecientes, y acelerantes y decelerantes.

• Lineales y no lineales. El tipo más simple y más familiar de transformación de datos es aquella que implica un cambio de escala, un cambio de origen o ambos simultáneamente (Emerson, 1991). Por ejemplo, cambiar de grados Celsius a grados Fahrenheit combina ambos:

$$C = \frac{5}{9}(F - 32)$$

Esta transformación está representada gráficamente en la figura 7.6.

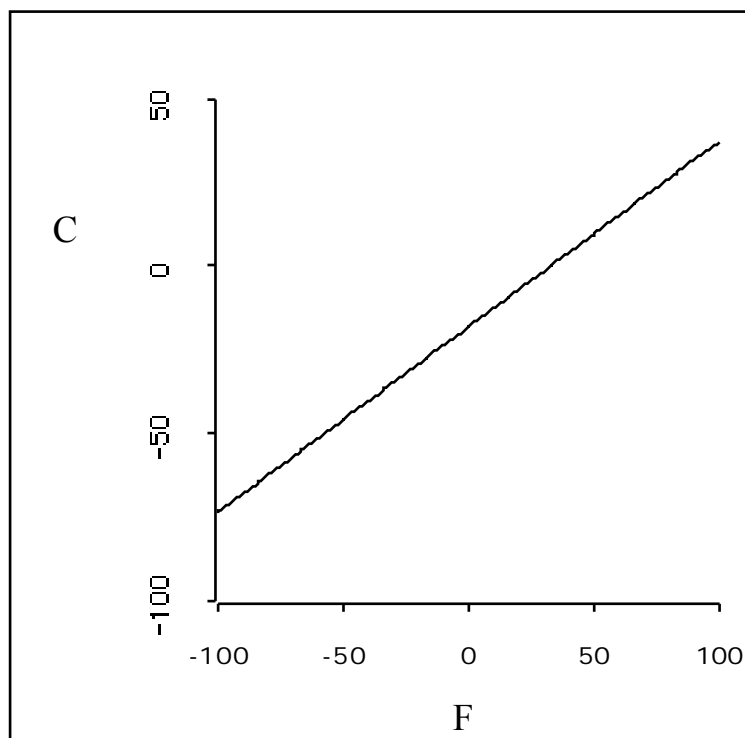


Figura 7.6: Transformación de grados Fahrenheit en grados Celsius

En este tipo de situaciones la transformación no supone un gran cambio desde el punto de vista de la interpretación o del significado de los valores. Puede utilizarse para, por ejemplo, eliminar puntos decimales o establecer un punto inicial más adecuado (por ejemplo, si existe un salario mínimo, este puede restarse de los salarios de un grupo de

trabajadores para hacer coincidir aquel con el cero) o cambiar la escala (como en el ejemplo anterior).

Una transformación lineal que implica un cambio en la interpretación de la escala es la que invierte los valores. La siguiente fórmula hace esto exactamente. Una representación gráfica de esta transformación es la de la figura 7.7.

$$y_i = (\max(X) - x_i + \min(X))$$

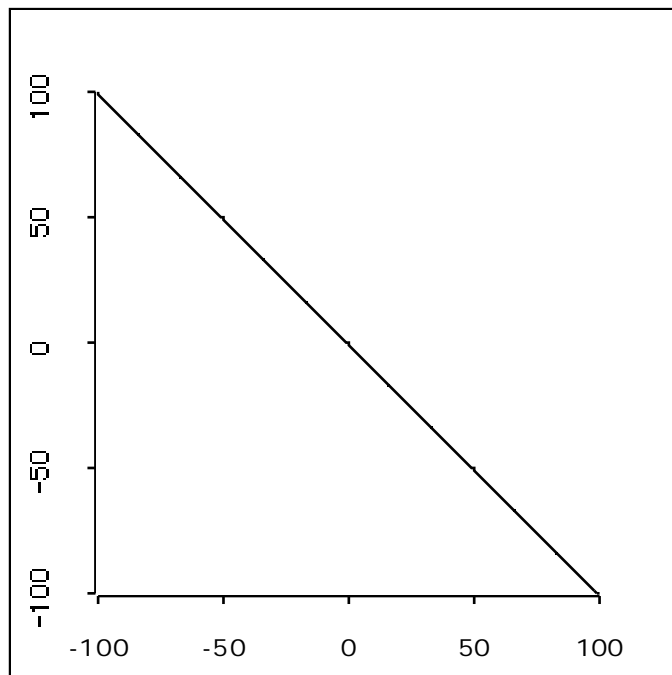


Figura 7.7: Transformación lineal inversa

Esta transformación es muy común en la evaluación de cuestionarios en los que ciertas preguntas se refieren a los mismos contenidos pero preguntados de tal modo que se oponen en sus valoraciones. El investigador puede preferir invertir aquellas preguntas que estén en sentidos opuestos para facilitar la interpretabilidad conjunta de todas ellas.

Una transformación no lineal modifica la variable de modo diferente según el valor considerado. Ambos gráficos en la figura 7.8 son transformaciones no-lineales. En ambos podemos ver que la pendiente de la curva varía en función de los valores considerados. En el de la izquierda por ejemplo, la pendiente es mayor al principio y luego disminuye.

- Transformaciones monotónicas y no monotónicas. Una transformación monotónica modifica los valores de la variable considerada siempre en la misma

dirección. Es decir, respeta el orden entre los valores. Una transformación no monotónica cambia este orden. Los dos primeros gráficos de la figura 7.8 corresponden a transformaciones monotónicas. La de la izquierda sería creciente y la de la derecha sería decreciente. El tercer gráfico corresponde a una transformación monotónica. Este tipo de transformaciones resultan poco interesantes en líneas generales desde un punto de vista estadístico.

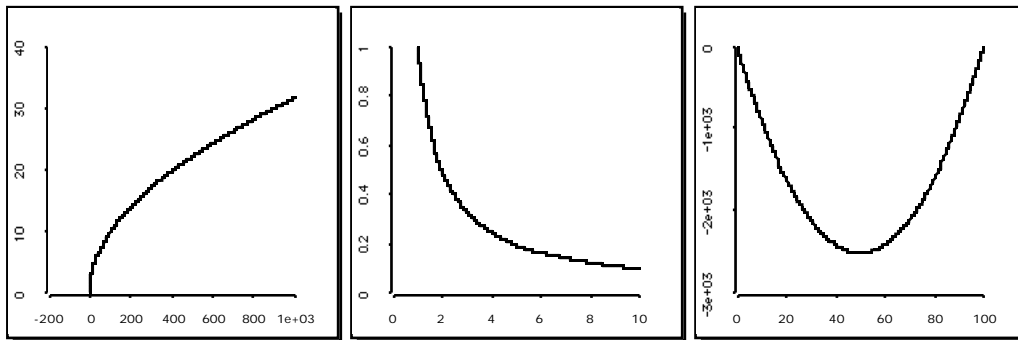


Figura 7.8: Transformaciones monotónicas y no monotónica

En el lado izquierdo se representa en primer lugar la transformación  $\sqrt{y}$ , en el segundo  $1/y$  y en tercero,  $y^2 - 100y$ . Generalmente, existe un problema con las dos primeras transformaciones con respecto a los valores negativos y el cero. Por ejemplo, no existe la raíz de valores negativos ni es posible dividir por cero. En la práctica, cuando es necesario realizar transformaciones de este tipo se realiza en primer lugar una transformación lineal del origen para eliminar los valores negativos y los ceros para luego realizar la transformación correspondiente.

- Transformaciones acelerantes y decelerantes. Una manera muy interesante de calificar las transformaciones anteriores es en términos de aceleración. La transformación  $\sqrt{y}$  (dibujada de nuevo en el lado izquierdo de la figura 7.9) es una transformación decelerante, mientras que la transformación  $y^2$  es una transformación acelerante. Esto implica que valores más grande en  $y$  se corresponde progresivamente en el primer caso con diferencias menores entre los valores transformados y en el segundo caso con diferencias mayores.



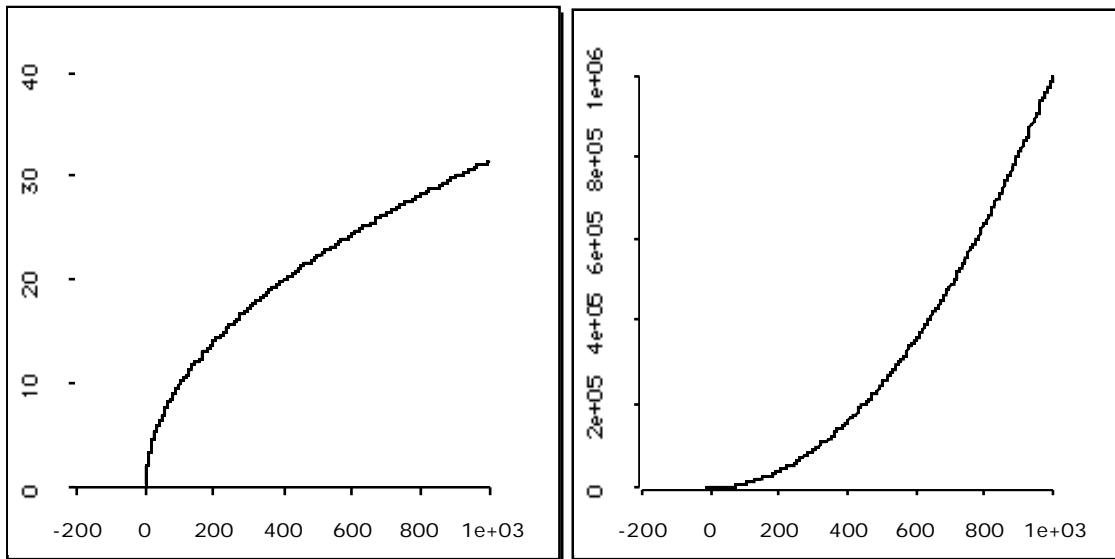


Figura 7.9 : Transformaciones monótona decelerante y acelerante.

La función  $y^3$  es simultáneamente acelerante y decelerante (con  $y=0$  indefinido) (figura 7.10).

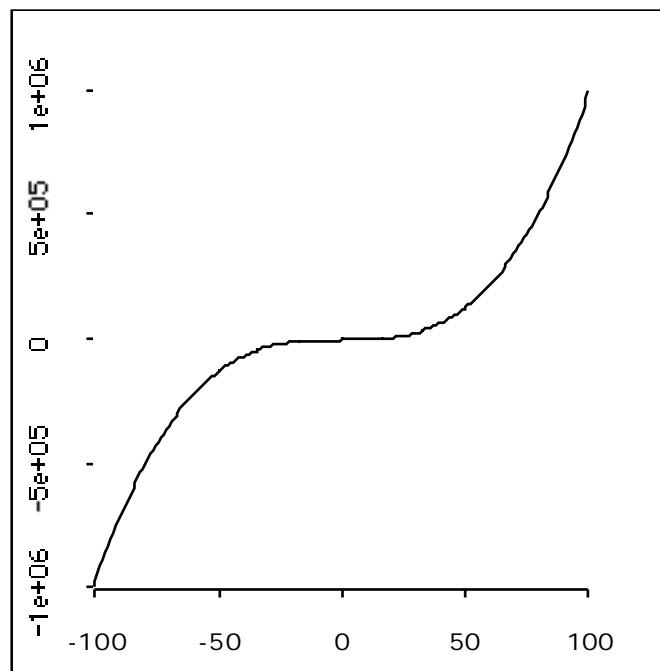


Figura 7.10: Función  $y^3$  como ejemplo de función acelerante y decelerante simultáneamente

### 7.3.2. La escalera de potencias

Un concepto que emerge de las distinciones anteriores es una clasificación de las transformaciones en función de su capacidad de doblar la curva de los gráficos mostrados anteriormente. En el extremo, las transformaciones lineales tienen una potencia nula puesto que el resultado es una línea recta. Otras transformaciones pueden ser más o menos acelerantes o decelerantes, dando lugar a diferencias en lo que hemos denominado potencias. Este concepto lleva a la denominada escalera de potencias que describiremos a continuación.

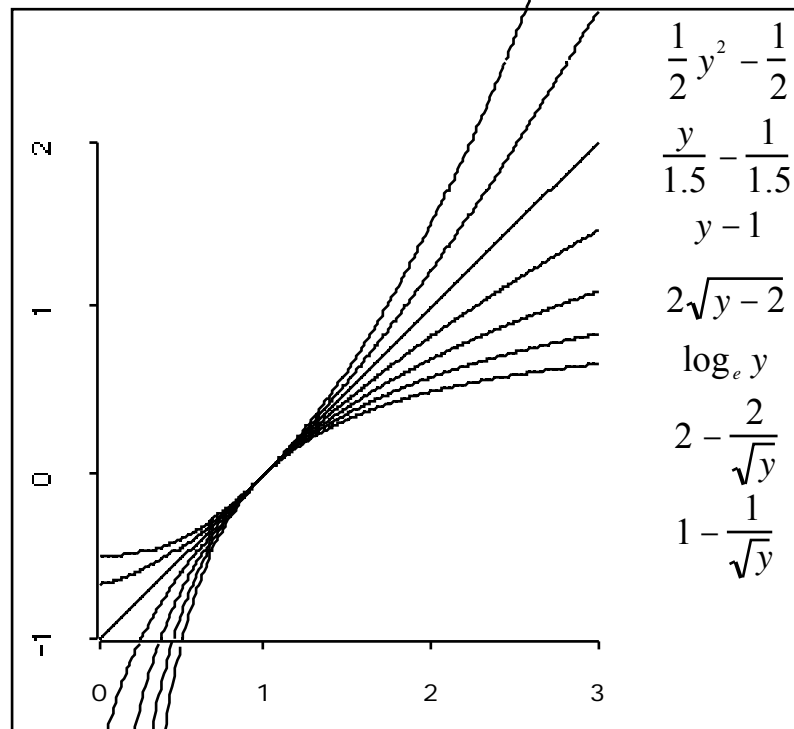


Figura 7.11: Escalera de potencias

En la escalera de potencias trataremos transformaciones de la siguiente forma.

$$\begin{aligned}
 & y^p \quad (p > 0) \\
 f(y) = & \log y \quad (p = 0) \\
 & -y^p \quad (p < 0)
 \end{aligned}$$

La fuerza y el tipo de transformación obtenidos al aplicar esta fórmula varía en función de  $p$ . Para valores de  $y$  superiores a 1, con  $p > 1$  la transformación es acelerante, con menor fuerza cuanto más cerca de 1. Con  $p = 1$  la transformación corresponde a la

identidad. Valores menores de  $p$  corresponden a transformaciones decelerantes, con más fuerza a medida que  $p$  es menor. Estos valores se invierten cuando  $y$  es menor de 1. Esto puede ser apreciado mejor si se normalizan los valores de  $f(y)$  del siguiente modo:

$$f(y) = \frac{y^p - 1}{p} \quad (p \neq 0)$$

$$\ln y \quad (p = 0)$$

La figura 7.11 muestra diferentes transformaciones para valores seleccionados de  $p$ . Como es posible ver, la transformación logarítmica, aunque no pertenece a la misma familia que el resto cae de un modo natural entre las correspondientes a  $p= 0.5$  y  $p= -0.5$ , llenando un hueco para el que no existe un valor apropiado dentro de las potencias. Este tipo de transformaciones son denominadas de Box-Cox (Atkinson and Cox, 1982, Emerson, 1991, Emerson and Stoto, 1983).

En la siguiente tabla se describen las transformaciones de Box-Cox en más detalle indicando cual es el efecto de cada una de ellas.

$p$	Función	Nombre	Efecto
2	$y^2$	Cuadrado	Aumenta más los valores grandes que los pequeños. Apropiado para datos con asimetría negativa.
1	$y$	Datos originales	No hay transformación
1/2	$\sqrt{y}$	Raíz Cuadrada	Disminuye más los valores grandes que los pequeños. Apropiaada para datos con asimetría positiva.
0	$\log(y)$	Logaritmo en base decimal (este valor sustituye a $X^0$ ya que este valor sería 1 siempre)	Lo mismo que el anterior pero su efecto es más extremo.
-1/2	$-1/\sqrt{y}$	Raíz recíproca (se le añade el signo menos para respetar el orden)	Lo mismo que el anterior pero su efecto es más extremo.
-1	$-1/y$	Recíproco (se le añade el signo menos para respetar el orden)	Lo mismo que el anterior pero su efecto es más extremo.

Tabla 7.12: Transformaciones del tipo Box-Cox

Emerett (1991) señala las siguientes razones para realizar transformaciones de datos:

- Facilitar una interpretación más natural de los datos.

- Conseguir simetría en unos datos.
- Conseguir mayor estabilidad de la varianza a través de varios grupos de datos.
- Conseguir una relación lineal entre dos variables.
- Simplificar la estructura de una tabla de dos dimensiones o superior de tal modo que un modelo simple aditivo sea capaz de ayudarnos a entender las características de los datos.

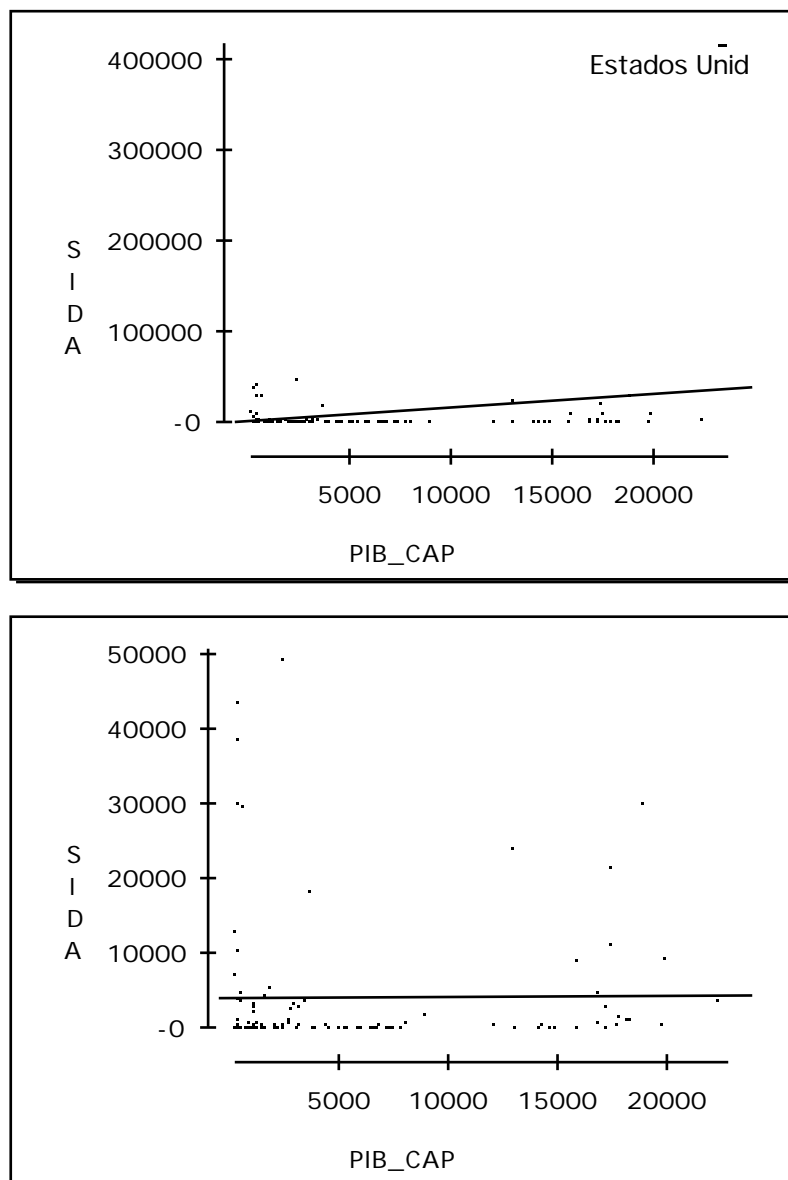


Figura 7.13: Diagramas de dispersión del PIB de varios países y el número de casos de SIDA antes y después de eliminar el dato de Estados Unidos

Las transformaciones de Box-Cox son en general guiadas por el segundo objetivo, conseguir mayor simetría en los datos. No obstante, resulta habitual que realizar una transformación en relación con un objetivo primario produzca de manera lateral beneficios con respecto a los otros objetivos. En la figura 7.13 es posible ver un ejemplo

En el gráfico se muestra la relación entre el logaritmo del Producto Interior Bruto Per Capita en países del mundo (datos tomados de SPSS) y el número de casos de SIDA de su población. El gráfico es bastante confuso (la mayoría de los datos se concentran en la parte inferior) y la relación lineal aparece poco clara y contraintuitiva (mayor logaritmo del producto interior bruto parece relacionarse con mayor número de casos de SIDA). Esta confusión parece deberse ante todo a Estados Unidos, señalado en la parte de arriba, el cual combina altos valores de producto interior bruto junto con un gran número de casos de SIDA. Eliminar Estados Unidos del gráfico no mejora su aspecto ya que parece haber ciertos países con bajas rentas per capita e inusualmente grandes niveles de SIDA. En este caso realizar una transformación de la variable número de casos de SIDA puede ayudarnos a hacer más clara esta relación. En concreto, nuestro objetivo es lograr una mayor simetría en la variable número de casos de SIDA ya que esto anticipamos que tendrá como consecuencia hacer más clara la relación mostrada en el diagrama de dispersión. Un histograma de la variable SIDA puede verse en la figura 7.14.

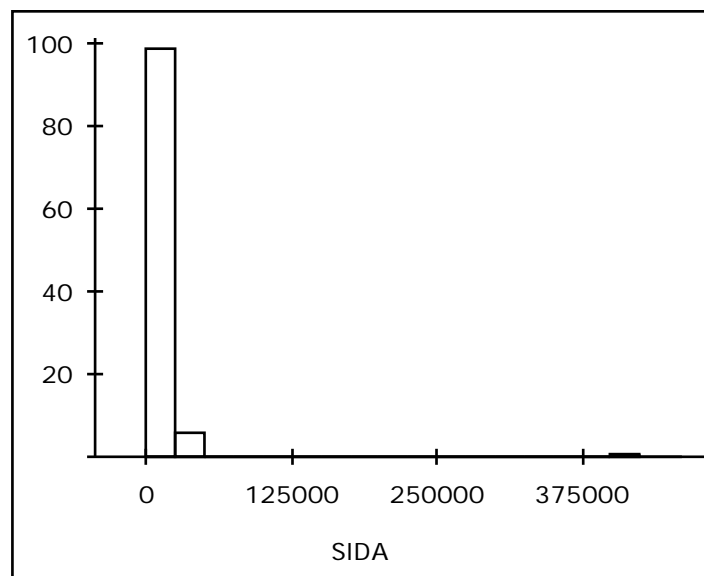


Figura 7.14: Casos de SIDA en países en el mundo

En la figura 7.15 se muestran los efectos de aplicar las transformaciones de la escalera de potencias sobre este histograma. Podemos ver que  $p > 0$  logran mejorar mucho la distribución inicial. Pero es  $p = 0$  (la transformación logarítmica) la que produce un

histograma bastante simétrico. Las transformaciones con  $p > 2$  podemos ver que invierten el histograma, dando lugar a asimetría positiva (es decir generando el problema contrario).

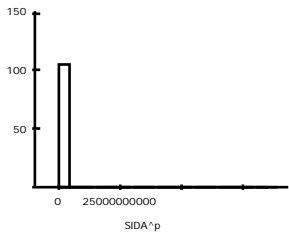
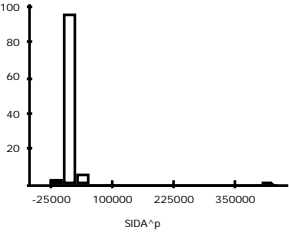
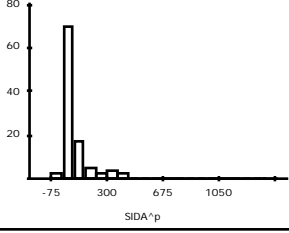
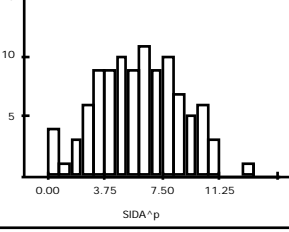
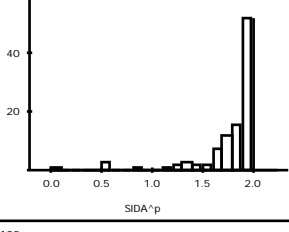
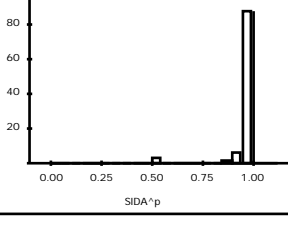
$p$	Función	Histograma
2	$y^2$	
1	$y$	
1/2	$\sqrt{y}$	
0	$\log(y)$	
-1/2	$-1/\sqrt{y}$	
-1	$-1/y$	

Figura 7.15: Efecto de las transformaciones de la escala de potencias.

La transformación correspondiente al logaritmo del SIDA produce simetría en histograma. El diagrama de dispersión de la variable transformada v. PIB\_CAP aparece más claramente como positiv ahora (figura 7.16).

Como es posible ver, una transformación que en principio estaba dirigida a producir mayor simetría en la variable modificada ha tenido como consecuencia hacer más clara una relación lineal (aunque todavía existe una considerable heteroscedasticidad). La transformación más adecuada ha sido obtenida mediante técnicas de visualización gráfica (Tierney, 1989, Velleman, 1995, Young, 1996) aunque métodos de máxima verosimilitud también están disponibles (Bozdogan and Ramírez, 1988).

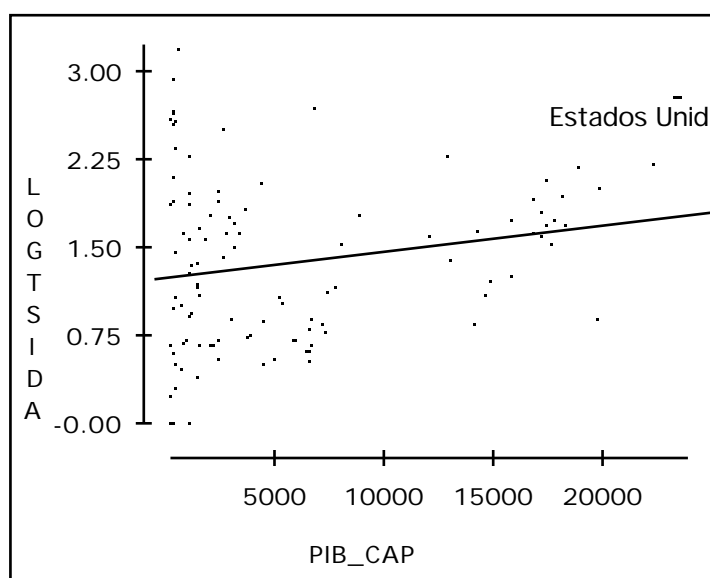


Figura 7.16: Diagrama de dispersión del logaritmo de SIDA y el PIB

No obstante, es posible objetar que esta transformación ha producido una variable con una interpretación más confusa. Esto puede considerarse un problema, así como el hecho que nuestras transformaciones han alterado la escala original de los datos de una manera que puede despertar sospechas de "legitimidad". Estas dos cuestiones serán tratadas al final de esta sección, después de haber discutido otras familias de transformaciones más especializadas.

### 7.3.3. Otras transformaciones

En ocasiones los datos son proporcionados en proporciones o porcentajes. Estos valores pueden beneficiarse de una transformación que aumente las diferencias en los extremos. Esto se justifica porque los cambios en los lados de la escala tienen más importancia que los del centro. Tomaremos como ejemplo el nivel de alfabetización en los

países del mundo de la fuente anteriormente mencionada. Estos datos están en proporciones de individuos alfabetizados en la población y son representados gráficamente en la parte inferior. Podemos ver que un reducido grupo de países tiene un nivel de alfabetización del 100% y que el número de países con niveles de alfabetización cercanos a 100 es muy alto. Podemos pensar que pasar de la situación en la que la gran mayoría de los individuos está alfabetizado a una en la que *toda* la población lo esté es un empeño muy difícil y que la diferencia numérica en la escala (1 o 2) no lo representa adecuadamente. Si comparamos este logro, obtener el 100% de la población alfabetizada para un país con un 97 o 98 por ciento de alfabetización, con un aumento similar en el centro de la distribución, pongamos de 66% al 68%, comprenderemos que el primero es mucho más difícil que el segundo, ya que, para alcanzarlo, será necesario poner en funcionamiento mecanismos especiales, mucho más costosos, que los requeridos para lograr el segundo cambio. Por ello, una transformación que aumente las distancias entre los datos en los extremos y deje relativamente poco modificados los situados en el centro parece interesante. En el caso de tratar con proporciones, esto nos proporciona valores que no se hallan limitados por los extremos sino que, dependiendo de la transformación utilizada, pueden variar entre - y + .

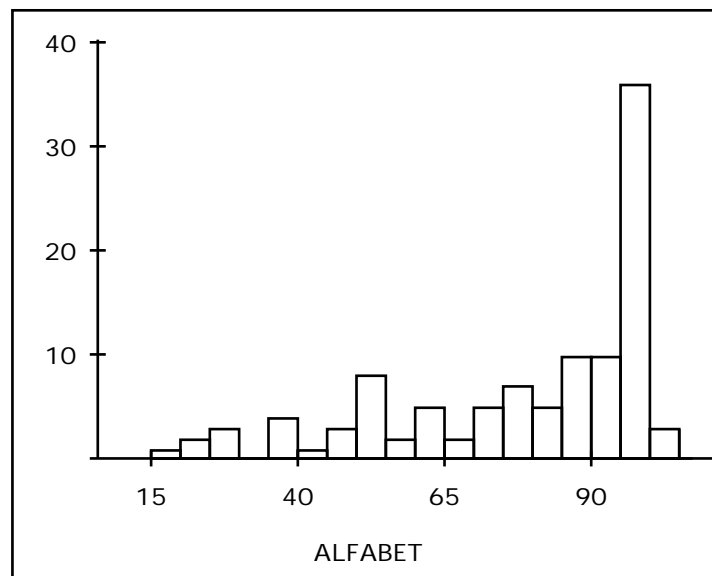


Figura 7.17: Histograma de la proporción de sujetos alfabetizados en países del mundo.

Una familia de transformaciones que actúa de este modo es la denominada *lambda* de Tukey. Esta transformación toma la siguiente forma general:



$$f(y) = \frac{p - (1-p)}{2} \frac{1}{\ln(p/1-p)} \quad ( > 0 )$$

$$\ln(p/1-p) \quad ( = 0 )$$

Aplicando diversos valores de se obtiene las siguientes funciones. De entre ellas, la primera, segunda y última son exactas. Las otras dos son aproximaciones muy cercanas (Velleman, 1995).

	<b>Función</b>	<b>Nombre</b>
1	$p-(1-p)$	<i>Pluralidad</i>
0.5	$\frac{\sqrt{p} - \sqrt{1-p}}{1/2} \frac{1}{\sqrt{2}}$	<i>Raíz cuadrada "doblada"</i>
0.41	$2\arcsin(\sqrt{p}) - \frac{2}{2}$	<i>Arcoseno</i>
0.14	$\frac{p^{0.14} - (1-p)^{0.14}}{0.14} \frac{1}{2^{0.14}}$	<i>Probit o Inversa de la transformación gaussiana</i>
0	$\ln(p/1-p)$	<i>Logit</i>

Tabla 7.18: Funciones para el cálculo de la lambda de Tukey

Velleman (1995) también recomienda realizar la siguiente transformación cuando los datos son jerarquías. Hacer:

$$p = \frac{i - 1/3}{n + 1/3}$$

En donde *i* son los rangos y *n* es el valor jerárquico superior. A continuación se debería aplicar las transformaciones en la lambda de Tukey. Esto permite "extender" los extremos un poco, acercándolos a lo que los valores originales subyacentes valdrían (o más bien hacia una versión de ellos antes de haber producido una transformación en busca de la simetría).

#### 7.3.4. La justificación de las reexpresiones

Después de haber examinado con detalle el tipo de transformaciones a los que nos referimos con el concepto de reexpresión resulta conveniente examinar con más profundidad los argumentos que justifican su aplicación. Hasta ahora, las únicas justificaciones aportadas son:

- Las escalas en que los datos nos son aportados son hasta cierto punto arbitrarias por lo que resulta admisible realizar modificaciones en ellas.
- Desde el punto de vista del análisis, aunque este punto no será desarrollado en detalle aquí, las transformaciones pueden hacer más adecuados nuestros datos.

La siguiente cita de Mosteller y Tukey (1977, p.89) explicita ambos argumentos:

*Los números son registrados y comunicados principalmente en formas que reflejan hábito o comodidad antes que en aquellas que son más adecuadas para el análisis. Como resultado, a menudo necesitamos re-expresar los datos antes de analizarlos.*

Esta visión se opone a la clásica de Stevens según la cual las escalas de medida reflejan propiedades de los datos en tal manera que sólo ciertas transformaciones son admisibles a menos que queramos "degradar" el nivel al que pertenecen. Como un recuerdo, las escalas propuestas por Stevens junto con las transformaciones admisibles para cada una de ellas son las de la tabla 7.19.

Esta clasificación de escalas de medida, en opinión de Velleman y Wilkinson (1993), induce a prácticas equivocadas. Algunas de las críticas que exponen son las siguientes:

- La clasificación de Stevens prohíbe ciertas transformaciones para cierto tipo de datos. Por ejemplo, las transformaciones incluidas dentro de la escalera de potencias son de tipo monótonico pero no lineal, por lo que su utilización (en caso que quisiéramos mantener el nivel de medida superior y no nos conformáramos con una reducción a nivel ordinal) sería inadecuada. Es más, puesto que el efecto de por ejemplo una transformación logarítmica es más acentuado en los extremos de los datos y es más cercano a una transformación lineal en el centro, desde el punto de vista de las escalas de Stevens estamos confundiendo las propiedades de nuestros datos. Para esta tipología, conceptos tales como linealidad, homoscedasticidad, aditividad o simetría no parecen tener significado.

<b>Escala</b>	<b>Transformaciones admisibles</b>
<ul style="list-style-type: none"> <li>• <i>Nominal: Son escalas que admiten solamente identificadores que no necesitan ni siquiera ser numéricos</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>Cualquier operación que no confunda o combine identidades.</i></li> </ul>
<ul style="list-style-type: none"> <li>• <i>Ordinal: Son escalas que transmiten información únicamente acerca del orden entre sus valores.</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>Sólo transformaciones monotónicas son admisibles (es decir, que respeten el orden entre sus valores)</i></li> </ul>
<ul style="list-style-type: none"> <li>• <i>Intervalo: En estas escalas la información hace referencia al orden entre los valores y a la distancia entre ellos (distancias iguales entre valores de la escala corresponden a distancias iguales entre los objetos a los que son asignados)</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>Las transformaciones admisibles son las que mantienen las diferencias relativas. En este caso por ejemplo, las transformaciones lineales (en las que añadimos un número a todos los valores de la escala o los multiplicamos por un mismo valor) son admisibles. No obstante, los logaritmos estarían excluidos.</i></li> </ul>
<ul style="list-style-type: none"> <li>• <i>Razón: Mantienen información acerca de las distancias y existe un cero absoluto.</i></li> </ul>	<ul style="list-style-type: none"> <li>• <i>Es posible multiplicar los datos por un valor, pero no podemos sumarles una constante (ya que eso modificaría el valor de cero).</i></li> </ul>

*Tabla 7.19: Escalas de medida de Stevens y transformaciones admisibles*

• Un buen análisis de datos no asume tipos de datos, sino que los deduce a partir del análisis. Un número de identificación de los casos puede pertenecer a una escala nominal o, si es asignado secuencialmente, transmitir un cierto tipo de información ordinal (por ejemplo, cuando este número indica el número de orden de una operación quirúrgica realizada por un equipo médico. Valores bajos indicarían menor experiencia y posiblemente peores resultados que valores más altos). El analista cuidadoso debería comprobar este extremo.

• Unos datos pueden ser considerados pertenecientes a tipos de escalas diferentes según la pregunta que estemos interesados en contestar. Velleman y Wilkinson (1993) ponen como ejemplo el número de cilindros en datos acerca de coches. Si los valores son

4, 6 o 8, estos pueden ser considerados pertenecientes a una escala nominal, ordinal o de intervalo según nuestros intereses.

- Existen muchos tipos de datos que no ajustan en las escalas de Stevens. Por ejemplo, las proporciones son valores limitados entre 0 y 1 y cualquier transformación tal y como multiplicar por una constante altera completamente su naturaleza.

- Existen procedimientos estadísticos que no pueden ser clasificados según los criterios de Stevens. Las medias recortadas por ejemplo tratan los extremos de los datos como si pertenecieran a un tipo de escala diferente del que pertenecen los valores centrales. Cowles (1989) se expresa del mismo modo.

- Los tipos de escala no son categorías precisas. Muchos datos reales no encajan en una categoría con claridad. Para Stevens, esto se solucionaría degradando los datos a la categoría inferior, lo cual tiene como consecuencia el tener que recurrir de modo excesivo a métodos no paramétricos.

La cuestión subyacente fundamental a estas críticas parece ser la consideración de la Teoría Axiomática de la Medida como una disciplina de tipo matemática, sujeta sólo a constricciones de coherencia interna y rigor formal antes que a contrastaciones con la experiencia real. Desde ese punto de vista, una clasificación que parta de los datos tal y como se producen puede ser más interesante al no limitar las situaciones que la Ciencia, que acepta la experiencia como prueba última de la validez de nuestros conocimientos, puede examinar.

Tukey (1977) propone una clasificación de los datos diseñada para servir como guía para transformar valores de datos, a menudo de un tipo a otro. Esta es:

- Nombres: Categorías de variables nominales. No existe ningún orden entre ellos.
- Grados: Categorías con cierto orden entre ellas pero que no ofrecen una unidad de medida claramente determinada. Escalas que ofrecen a los sujetos 5 o 7 opciones de bajo a alto son ejemplos de este tipo.
- Rangos o jerarquías: Valores que indican el orden pero no el valor de los casos.
- Recuentos. Son números enteros que indican cuantos elementos hay de una determinada categoría. Por ejemplo, hay 25 hombres y 12 mujeres. Hay 10 bolas rojas, 5 negras y 3 verdes, etc.

- Fracciones de recuentos. Son cocientes con una base fija. Por ejemplo, hay 25 hombres de un total de 37 individuos por lo que  $25/37$  son hombres y  $12/37$  son mujeres. Algunas de las fracciones de recuentos más usuales son las proporciones (base =1) y los tantos por ciento (base=100).
- Cantidades. Este es el tipo más habitual de datos. No pueden ser negativos y como ejemplo tenemos cantidades de dinero, altura de objetos, distancias etc.
- Balances. Pueden ser positivos y negativos y son la forma más general de datos, aunque la menos común. Por ejemplo, el saldo de una cuenta corriente, donde los valores en rojo indican valores negativos. A menudo provienen de la diferencia entre dos cantidades.

Los recuentos y las cantidades a menudo pueden ser transformados mediante alguna de las opciones en la escalera de potencias. Los balances a menudo no pueden ser mejorados mediante una transformación, pero sí las cantidades de las que provienen. Para las fracciones de recuentos resulta apropiada la *lambda* de Tukey. Las jerarquías también pueden ser tratadas tal y como se expuso anteriormente.

Finalmente, es necesario advertir que, tal y como señalan los mismos Velleman y Wilkinson (1993), resultaría incorrecto considerar el concepto de escalas de medida como inútil. Por el contrario, su valor es indudable y la discusión anterior no habría sido posible sin este concepto. Sin embargo, lo equivocado es limitar el análisis de los datos, en este caso las transformaciones admisibles, a la escala de medida que aparentemente tenemos. Por el contrario, resulta mucho más razonable que, cuando finalmente se ha alcanzado una conclusión de algún tipo, entonces comprobar si resulta razonable asumir que los datos utilizados para llegar a esa conclusión satisfacen los niveles de medida.

Sarle (1995) considera que las críticas de Velleman y Wilkinson (1993) contra la teoría de la medida son exageradas e incorrectas. Para él, considerar las cuestiones acerca de las escalas de medida es necesario si lo que uno quiere es hacer inferencias acerca de lo que subyace a aquello que se ha medido y no limitarse simplemente a los números que uno tiene ante sí. Por otro lado, la teoría de la medida no dicta que estadísticos son apropiados para unos datos a un nivel específico de medida, pero el método estadístico debería producir resultados invariantes bajo las transformaciones admisibles a ese nivel de medida. Por ejemplo, el coeficiente de variación no es invariante a transformaciones monótonicas en datos a nivel de intervalo y no produce resultados equivalentes ante una simple transformación de, por ejemplo, la temperatura en de grados Celsius a Fahrenheit.

En definitiva, las transformaciones nos pueden ayudar a hacer un análisis estadístico más apropiado bajo el punto de vista de los supuestos que subyacen a determinados modelos (p.e. linealidad, homoscedasticidad, etc.). No obstante, la modificación del significado de los datos debería ser tenido en cuenta, de tal modo que tiene que ser claro que las inferencias lo son acerca de las variables transformadas y no las originales.

## **7.4. El registro del proceso de manipulación de ficheros**

Uno de los problemas de la gestión de datos se produce cuando se realizan tareas de transformaciones para satisfacer objetivos concretos y se van produciendo nuevos conjuntos de datos que difieren del original. Esto puede llevar a confundir unos con otros y a realizar análisis sobre datos que no corresponden a lo deseado. Davidson (1996) ofrece una serie de reglas de sentido común con las que él se siente satisfecho. En nuestra opinión, cualquier conjunto de reglas establecidas explícitamente y desarrolladas por cada investigador puede ser suficiente.

Otra alternativa es utilizar un software diseñado para llevar registro de las transformaciones y análisis realizados. ViSta (1998) ofrece un espacio de trabajo diseñado para satisfacer este objetivo que puede verse en la figura 7.20. En él puede observarse que el archivo "CrimeRates" ha sido transformado por un lado en puntuaciones normales, lo cual ha producido como resultado un nuevo archivo que corresponde a éstas. A continuación se ha aplicado una ordenación de los datos en función de una variable. Por otro lado, vemos que el archivo original fue transformado aplicando la función logarítmica y que, a continuación, se aplicó un modelo de componentes principales.

Este tipo de representación puede ser de gran ayuda al analista (sobre todo si pudiera ser guardado en disco y posteriormente recuperado) ya que le permitiría saber en cada momento en qué estado se encuentra el archivo de datos que quiere utilizar y cual es la historia que le precede. Paquetes estadísticos tradicionales ofrecen un archivo de registro (log) que ofrece una información similar para cada sesión. Sin embargo, mantener ese registro para diferentes sesiones resulta complicado.

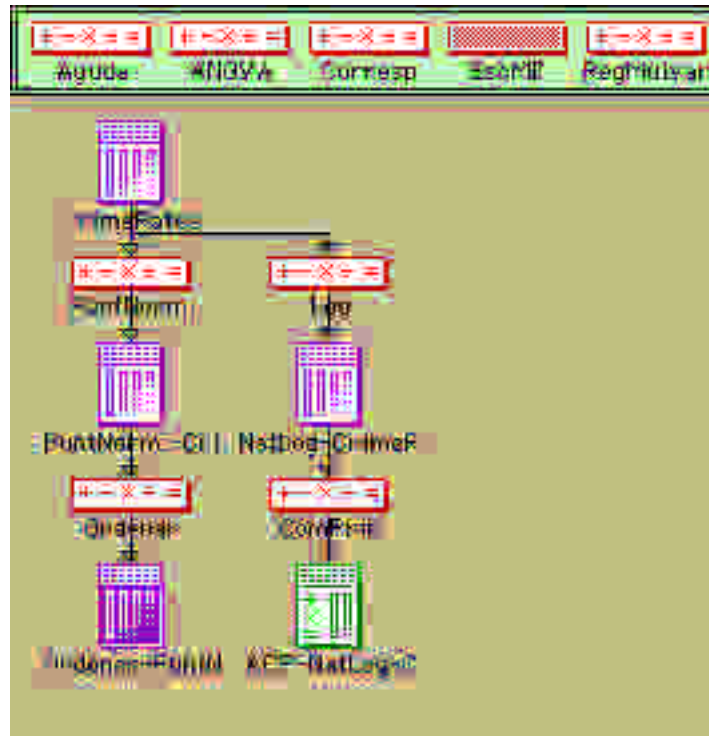


Figura 7.20: Espacio de trabajo de ViSta





# 8. Programación estadística

## 8.1. Introducción

A medida que se dominan mejor las herramientas básicas del proceso de datos los usuarios ven aparecer normalmente dos nuevas necesidades. En primer lugar, desean que ciertas tareas de índole repetitiva se puedan programar para ejecutarse de modo automático, y, en segundo lugar, les gustaría poder aumentar las capacidades de tratamiento de datos de tal modo que sea posible realizar tareas que antes eran muy difíciles o imposibles.

Normalmente, para cubrir la primera de las necesidades la solución es acudir a lenguajes de programación que proporcionan los paquetes estadísticos, programas de bases de datos u otras aplicaciones. Estos lenguajes suelen permitir poner los comandos que les son propios en sucesión, de tal modo que sea posible ejecutarlos uno detrás de otro, sin necesidad de supervisión del operador. De hecho, los sistemas operativos permite la interconexión de programas entre sí de modo que es posible utilizar ciertos lenguajes para transferir información entre ellos para aprovechar sus potencialidades diferentes. Un ejemplo puede ser el siguiente. Supongamos una base de datos que incluye información acerca de los alumnos de enseñanza secundaria en los institutos de una ciudad. Su objetivo es proporcionar información descriptiva y comparaciones

estadísticas acerca de estos institutos con una periodicidad trimestral. Esto implica extraer la información de la base de datos y exportarla al formato de un paquete estadístico, importarlo desde éste, llevar a cabo los cálculos correspondientes, y no menos complicado, imprimir los resultados con un aspecto apropiado. Puesto que este informe tendrá un esquema similar cada trimestre, el encargado de realizarlo pronto buscará la manera de automatizar el proceso para hacerlo más simple.

La segunda necesidad corresponde al desarrollo de nuevos procedimientos estadísticos, o, al menos, a modificaciones de los ya creados que permitan añadir ciertos cálculos en los que estemos interesados. Por ejemplo, supongamos que en cierto paquete estadístico no está disponible el coeficiente de fiabilidad entre ítems pares e impares en un test. El usuario puede conocer otros programas que sí lo calculan, pero el esfuerzo necesario de traducir los datos puede ser excesivo comparado con el de simplemente programar este coeficiente. La dificultad de esta programación dependerá de las características del sistema que estemos utilizando ya que en ocasiones puede ser impracticable.

Describiremos tres grupos de herramientas en relación con la programación para el tratamiento de datos estadísticos. En primer lugar, lenguajes de programación de paquetes estadísticos. En segundo, entornos de programación estadística y, finalmente, lenguajes de programación propiamente dichos. Con ello dejamos fuera muchos posibles lenguajes o métodos de programar/automatizar tareas, centrándonos en aquellos que, en principio, tienen más importancia para el tratamiento de datos estadísticos.

Existe cierto grado de solapamiento entre estas herramientas. Todas ellas pueden ser utilizadas para llevar a cabo tareas semejantes. No obstante, las diferencias aparecen muy rápidamente si atendemos a la facilidad con que éstas pueden ser llevadas a cabo. Por ejemplo, si el usuario desea utilizar un procedimiento bien conocido y disponible en paquetes estadísticos, pero de manera repetitiva, lo más aconsejable será utilizar el lenguaje de éstos. Por ejemplo, SPSS o SAS ofrecen facilidades para este tipo de tareas. Por el contrario, si se desea desarrollar nuevos procedimientos estadísticos lo más apropiado puede ser un entorno de programación tal y como S-Plus o Lisp-Stat. En cambio, si el desarrollo estadístico es lo suficientemente interesante como para invertir esfuerzo en hacerlo lo más eficiente posible puede ser interesante desarrollarlo en un lenguaje de programación de corte clásico que disponga de un compilador. Veamos a continuación algunas de las características que los distinguen.

a) Lenguajes de paquetes estadísticos. Existen una serie de paquetes estadísticos que podemos denominar como tradicionales o clásicos. Estos corresponden a productos

que han logrado mantenerse en el mercado durante bastantes años (a menudo desde los años 60-70) y que han ido extendiendo sus capacidades, los sistemas operativos en los que funcionan, la documentación disponible y los usuarios durante todo este tiempo. Los dos ejemplos más usualmente nombrados son SPSS (Statistical Package for the Social Sciences-Paquete Estadístico para las Ciencias Sociales) y SAS. Otros paquetes que tuvieron una carrera comercial relativamente amplia pero que actualmente parecen haber terminado con ella son Systat (the System for Statistics-Sistema para Estadística) y BMDP (Biomedical Data Package-Paquete de Datos Biomédicos).

Estos programas comparten el haber sido diseñados con un modo de utilización en diferido ('batch'). Esto significa que los usuarios escribían los comandos que deseaban ejecutar en el ordenador y los guardaban en un archivo. Posteriormente, cuando el ordenador en el que ese programa iba a ser ejecutado se encontraba "libre" el programa era lanzado. Los resultados normalmente eran impresos en papel inmediatamente y consultados para determinar si eran los esperados. En caso de producirse errores se repetía el proceso. En aquellos tiempos utilizar un paquete estadístico no difería prácticamente nada de lo que actualmente denominamos programar un paquete estadístico. La bajada de precios y la mejora de las capacidades de los ordenadores personales en cuanto a cálculo y posibilidades gráficas pronto condujo a que cada usuario dispusiera de su propio ordenador. Esto supuso un cambio en la forma en que este tipo de programas estaban concebidos, en la dirección de dotarlos de una mayor interactividad. Así, usar un paquete estadístico se convirtió en una tarea de ejecutar un comando, obtener un resultado y pasar al siguiente comando en caso de estar satisfecho con lo obtenido. Esto llevó a que ciertas características de los paquetes estadísticos se tornaran obsoletas. Por ejemplo, generar enormes cantidades de resultados con cada análisis es innecesario puesto que el usuario puede pedir aquello que desee de modo interactivo (Tierney, 1990). También, las nuevas capacidades gráficas de los ordenadores permitieron la utilización de técnicas de análisis poco prácticas hasta el momento (Chambers, et al., 1983).

No obstante, a pesar que estos paquetes estadísticos añadieron la posibilidad de utilizar menús y otros elementos de interacción como cuadros de diálogo, ventanas, etc., los antiguos lenguajes de programación permanecieron, aunque su función se vio reducida. En la actualidad, estos lenguajes son utilizados en primer lugar para llevar registro de lo realizado, y, en segundo, continúan siendo utilizados para formar paquetes pero sólo cuando estos van a ser utilizados de modo repetitivo y resulta muy ventajoso hacerlo.

b) Entornos de programación estadística. Este tipo de herramientas se centran en desarrollar nuevos procedimientos estadísticos. Así, poseen lenguajes que incluyen operaciones de álgebra matricial, transformaciones de datos, representación gráfica, etc., así como funciones típica mente informáticas (estructuras condicionales, bucles, ...).

Así, a través de estas operaciones, algunos de ellos incorporan procedimientos estadísticos que rivalizan en cuanto a capacidades con los de paquetes estadísticos considerados clásicos, con la ventaja adicional de una potencialidad y flexibilidad con la que aumentar sus capacidades.

Ahora bien, algunas de ellas incorporan ya los más clásicos, de tal modo que sus capacidades a menudo se asemejan, sin ningún esfuerzo por parte del usuario, a las de los paquetes estadísticos, con la ventaja de esa mayor potencialidad y flexibilidad disponible para aumentar sus capacidades.

Estos lenguajes poseen una historia relativamente corta, cuyo inicio podría bien situarse en como máximo hace una o dos décadas. Un acontecimiento usualmente nombrado como clave es la definición del lenguaje S (Becker and Chambers, 1984, Becker, et al., 1988) el cual ha influido en los dos sistemas en que nos centraremos en este texto: S-Plus (Data Analysis Products Division, 1997) y Lisp-Stat (Tierney, 1990).

Tierney (1990) señala que existen dos posibles aproximaciones a la creación de un lenguaje de programación estadística. Desarrollar un nuevo lenguaje desde el principio o bien extender un lenguaje de programación convencional para incorporar funciones desde el punto de vista de lo que un estadístico desea. S-Plus toma el primer camino, mientras que Lisp-Stat es una modificación de Lisp que incorpora funciones estadísticas. No obstante, en realidad las diferencias entre ambos no son tan importantes y ambos parecen haber tomado prestadas características entre sí.

Del mismo modo, APL2STAT parte del lenguaje APL y toma un camino muy similar a Lisp-Stat (Stine and Fox, 1997b)., aunque parece haber recibido menos atención. Otros lenguajes que han sido considerados para cubrir esta función como GAUSS y Mathematica proporcionan una buena base matemática pero parecen ser demasiado generales y menos cercanos a las tareas estadísticas.

c) Lenguajes de programación generales. Existe una gran variedad de lenguajes de programación. Lo siguiente es una revisión necesariamente rápida.

FORTTRAN es el lenguaje de programación más antiguo, sin tener en cuenta los de bajo nivel tipo ensamblador. Su objetivo de diseño fue facilitar la traducción de fórmulas

matemáticas al ordenador (FORmula TRANslator es el significado de sus siglas). En su tiempo este era el lenguaje por excelencia de la programación matemática y estadística pero su uso parece estar descendiendo. Entre sus inconvenientes parece encontrarse el no haberse adaptado a la tendencia hacia la programación orientada al objeto ocurrida en los últimos años aunque a su favor está la existencia de una gran cantidad de código para muchos problemas. Nacido en los años 50 su uso se convirtió en tan popular en los 60 que aparecieron una gran cantidad de dialectos en los años 60. Ello llevó a un esfuerzo de acuerdo entre diferentes vendedores que concluyó con Fortran 66, el primer lenguaje de programación que fue oficialmente estandarizado. En los años 70 se convirtió en el lenguaje más utilizado para propósitos no comerciales y un poco más tarde se desarrolló un nuevo standard, Fortran 77. En la actualidad este lenguaje puede considerarse que está pasado desfasado y en la actualidad está disponible Fortran 90. Ahora bien, dada la gran cantidad de programación disponible desde los años 60 escrito en versiones antiguas y que, dada la gran cantidad de experiencia con su funcionamiento, puede considerarse fiable y adecuado para sus objetivos, todas las nuevas versiones de Fortran se esfuerzan en hacer el código previo compatible con las nuevas. Además, este lenguaje ha sido optimizado de tal manera a lo largo de los años que el código obtenido es más rápido que el de otros lenguajes más modernos como Pascal o C. Además, se han intentado añadir características que el programador pueda utilizar en lugar de aquellas que ya son conocidas por sus posibles inconvenientes (la información aquí descrita ha sido tomada de <http://unics.rrzn.uni-hannover.de/rrzn/gehrke/HPFKurs/HTMLHPFCourseNotes/node3.html#SECTION01011000000000000000>).

Lisp es el segundo lenguaje en cuanto a edad en la actualidad (nació un año después de FORTRAN). Su origen fue facilitar la representación y manipulación de información simbólica. Common Lisp es una implementación de Lisp que se ha convertido prácticamente en un standard en la actualidad. Este lenguaje ha sido tradicionalmente considerado el más apropiado para problemas de inteligencia artificial y en general para problemas complejos, poco definidos, con estructuras de datos desconocidas antes del momento de ejecución y que pueden necesitar de cambios y modificaciones constantes. Así, puede verse como una herramienta apropiada para la programación experimental e incremental. En su contra está que toda su indefinición impide una adecuada optimización por lo que el código obtenido no resulta tan rápido como el obtenido con otros lenguajes a más bajo nivel.

Pascal es un lenguaje mucho más moderno que Fortran en el sentido que promueve buenas técnicas de programación. De hecho fue utilizado de manera habitual como primer lenguaje de programación en las universidades. En la actualidad C, el cual comparte características con Pascal, es mucho más popular. No obstante, al ser desarrollado por

programadores para programadores, muchas de sus características no son consideradas apropiadas para promover buen estilo por lo que, aunque es el lenguaje con mayor proyección profesional en la actualidad no suele ser visto como el apropiado para usuarios novatos. C está asociado al desarrollo del sistema operativo Unix, y en los últimos tiempos ha sido extendido con características de orientación a objetos (C++).

Otros lenguajes que merecen al menos ser citados son:

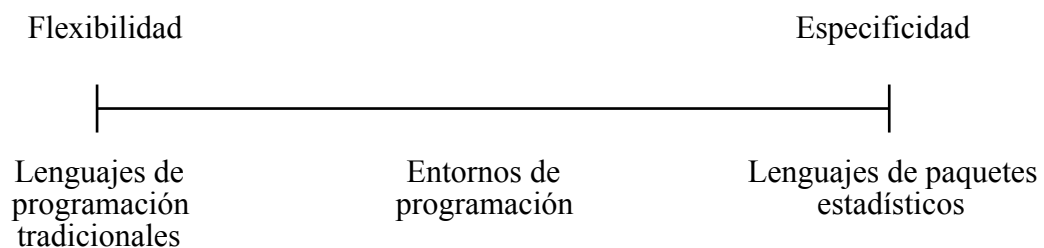
- Basic, un lenguaje diseñado para enseñar programación pero que debido a sus características de fácil utilización ha sido considerado como el más apropiado para usuarios no profesionales. En sus sucesivas reencarnaciones ha ido incorporando diversas características hasta la última, de tipo Visual, que parece haberle dado un empuje renovado.

- Java, un lenguaje diseñado para facilitar la portabilidad entre diversos ordenadores y sistemas operativos y cuyo desarrollo parece estar asociado al de la Internet. Lleno de promesas en un primero momento parece que problemas comerciales están deteniendo su crecimiento.

## 8.2. Comparación entre tipos de lenguajes

La cuestión que surge naturalmente después de la sección anterior es la de establecer qué diferencias existen entre los diferentes tipos de lenguajes considerados: lenguajes de paquetes estadísticos, entornos de programación y lenguajes de programación tradicionales.

En primer lugar está la cuestión de la *flexibilidad v. especificidad* del lenguaje. Si ponemos los tres tipos de lenguajes en un continuo en el que ambos conceptos se encuentran en los extremos obtenemos la representación de la figura 8.1:



*Figura 8.1: Flexibilidad v. especificidad*

Por flexibilidad entendemos el número de tareas diferentes que es posible programar mediante un lenguaje, así como la libertad disponible para enfocar los problemas. Por especificidad entendemos lo bien que cubren el tipo de tareas en las que estemos interesados, en este caso, las estadísticas. En general, los lenguajes de programación tradicionales son más flexibles pero mucho menos específicos. Por medio de ellos podrían programarse tanto las tareas estadísticas como otras muy diferentes de aquellas (sistemas operativos, procesadores de texto, etc.). Los lenguajes de paquetes estadísticos, en cambio, prácticamente sólo permiten programar tareas de tipo estadístico, y limitados a la forma en que el paquete concreto que estemos usando concibe estas tareas. Por ejemplo, el modelo de datos a utilizar o los tipos de gráficos disponibles no pueden ser alterados.

Los entornos de programación estadística quedan en un punto intermedio entre ambos ya que estos lenguajes son capaces de tener especificidad a la vez que tienen una gran flexibilidad. De este modo, tareas tan mundanas como importar un archivo con una estructura difícil puede ser realizado con facilidad.

Por otro lado, lo cierto es que los paquetes estadísticos llevan mucho tiempo en el mercado, y en la práctica, se han extendido tanto que prácticamente cubren la mayoría de las necesidades que podamos tener, a pesar de parecerlos peculiares o poco comunes. De este modo, utilizarlos nos evita reinventar soluciones a problemas ya bien resueltos por ellos.

Así pues, existen argumentos a favor de la flexibilidad y la especificidad, lo cual hace que coexistan todo tipo de herramientas.

Una segunda cuestión es la referida a lenguajes *interpretados* vs. *compilados*. En un lenguaje interpretado, el código que construimos se ejecuta instrucción a instrucción. Esto hace una aplicación informática denominada *interpretador* la cual deberemos estar haciendo funcionar para llevar a cabo esta ejecución. Puesto que el interpretador mira cada instrucción individualmente es incapaz de optimizar el código de modo global tal y como ocurre cuando es *compilado*. Un *compilador* es también una aplicación informática que es capaz de convertir un programa a un lenguaje muy cercano al del computador. En este proceso optimiza el código y permite que posteriormente sea ejecutado por sí mismo, sin la ayuda de otro programa.

Una consecuencia que tiene compilar código es que éste ya no es directamente accesible, puesto que la versión más cercana al computador no está hecha para ser entendida por la gente.

En general, los lenguajes de paquetes estadísticos suelen ser de tipo interpretado. Los entornos de programación estadística suelen tener un enfoque combinado, en el que ciertas partes son interpretadas y otras son compiladas, aunque su esquema general de funcionamiento se asemeja al de los interpretadores. En ellos, ciertas partes del programa pueden ser compiladas para mayor eficiencia y posteriormente llamadas cuando se necesite, tal y como si fueran operaciones externas. Los lenguajes de programación clásicos suelen estar orientados a la compilación (aunque hay excepciones).

La ventaja de los lenguajes interpretados es que permiten el desarrollo incremental entre diferentes usuarios. Así, alguien puede programar una serie de operaciones estadísticas y ofrecerlas públicamente (usualmente a través de Internet) y, más tarde, otros usuarios pueden tomarlas, utilizarlas, y añadir nuevas funciones para completar aspectos no considerados. Además, si tienen acceso al código pueden incluso mejorarlo en determinados aspectos, bien por sí mismos, o bien sugiriéndoselos al autor (de nuevo a través de Internet). Ello es todavía más fácil cuando el lenguaje que utilizamos facilita un estilo de *programación funcional y orientada al objeto* (conceptos que discutiremos más adelante).

Ya centrándonos en los lenguajes interpretados, resulta interesante mencionar algunas de las diferencias señaladas por Stine y Fox (1997) entre lenguajes de paquetes estadísticos y entornos de programación estadística. Estas son:

a) Programabilidad: Los paquetes estadísticos funcionan con procedimientos preprogramados por completo. Los entornos estadísticos permiten acceso a los bloques básicos con los que estos procedimientos están contruidos. Ello permite construir nuevos procedimientos estadísticos.

b) Ampliación. Construir un nuevo procedimiento estadístico con las herramientas proporcionadas por los paquetes estadísticos resulta bastante difícil. Por ello, esta tarea queda reservada a los desarrolladores "oficiales", generalmente los de la propia empresa que comercializa el producto,. En los entornos de programación la diferencia entre usuarios y desarrolladores desaparece y resulta fácil para los primeros añadir nuevos procedimientos estadísticos.

c) Modelo de datos flexible. Los paquetes estadísticos comunes sólo ofrecen conjuntos de datos rectangulares, de la forma de filas por columnas. Aunque los desarrolladores que producen esos paquetes tienen acceso a estructuras más complejas y flexibles, éstas no están disponibles para los usuarios, lo cual les limita mucho en cuanto a los posibles resultados que pueden obtener.



Además de las mencionadas por Stine y Fox es posible también tener otras en cuenta.

a) Calidad del software. Puesto que los usuarios son capaces de desarrollar y distribuir software existen menos garantías acerca de su precisión que cuando es una empresa, con una reputación a salvaguardar, lo que lo hace. En la práctica, esta objeción no parece una preocupación de importancia.

Por otro lado, el software puede adolecer de falta de eficiencia aunque sea correcto. De hecho, el manual del programador de S-Plus (Data Analysis Products Division, 1997) menciona que el criterio de la eficiencia no debería impedir al usuario centrarse en la oportunidad de la experimentación y el desarrollo rápido que esta herramienta ofrece. Ello puede llevar a que estos entornos pueden no ser los adecuados para llevar a cabo análisis de cierta envergadura, y sea preferible utilizar paquetes estadísticos convencionales.

b) Comandos de bajo nivel. Los comandos de bajo nivel de los entornos de programación estadística permiten realizar tareas que los paquetes estadísticos no podrían realizar, tal y como, por ejemplo, un problema tan sencillo como cambiar comas decimales (continente europeo) a puntos decimales (Inglaterra, Estados Unidos y en general países anglosajones).

### **8.3. Descripción de entornos de programación estadística.**

Los entornos de programación estadística merecen una descripción especial. En Stine y Fox (1997) se mencionan hasta siete sistemas que podrían encajar dentro de esta descripción pero la lista podría ser más amplia. A continuación describiremos brevemente los sistemas mencionados en este libro agrupándolos por bloques. Posteriormente nos centraremos en los dos sistemas (S-Plus y Lisp-Stat) que probablemente mejor encajan con nuestro de entorno de programación estadística.

Los bloques son los siguientes:

a) Lenguajes de manipulación simbólica apropiados fundamentalmente para matemáticas. El sistema descrito en Stine y Fox (1997) es Mathematica pero existen otros de similares características (Derive, Maple, MACSYMA).

b) Lenguajes de programación de paquetes estadísticos. Los dos citados en este libro son SAS y Stata.

c) Lenguajes de manipulación de matrices. GAUSS estaría dentro de esta categoría. Además, debido a que posee capacidades para calcular regresiones y otros comandos estadísticos, ha sido utilizado para hacer análisis estadísticos.

d) Entornos de programación estadístico. En este apartado estarían APL2STAT, Lisp-Stat y S-Plus. Los dos primeros comparten el ser lenguajes de programación clásicos con extensiones para incorporar tareas estadísticas, mientras que el tercero se origina como un nuevo lenguaje creado específicamente para esta tarea. Los dos últimos parecen haber sido los que han alcanzado más desarrollo y serán descritos a continuación.

*S-Plus* (<<http://www.mathsoft.com/>>) es la implementación comercial de un lenguaje denominado S (Becker and Chambers, 1984, Becker, et al., 1988) especialmente diseñado para la programación estadística. En realidad, la implementación comercial incorpora funciones y mejoras que no estaban disponibles en la formulación original de S, tal y como la orientación a objeto (Stine and Fox, 1997a). Este lenguaje ha adquirido tanta importancia que procedimientos estadísticos son proporcionados por medio de este lenguaje antes de programados por medio de un lenguaje tradicional. Existen implementaciones para el sistema operativo UNIX y Microsoft Windows. Entre sus inconvenientes suele señalarse que exige una inversión relativamente importante en hardware. Existe también una versión gratuita de S denominada R (<http://stat.auckland.ac.nz/rproj.html>).

*Lisp-Stat* (<<ftp://ftp.stat.umn.edu/pub/xlispstat/current>>) está basado en el lenguaje Lisp pero incluye muchos añadidos que forman parte del lenguaje S y han sido tomados de él. Otras operaciones simplemente han sido modificadas para hacerlas funcionar con vectores y matrices. Según su creador (Tierney, 1990) la motivación para el desarrollo de Lisp-Stat fue su interés en explorar el uso de gráficos dinámicos así como el tipo de interfaz de usuario más apropiado para interactuar con ellos. Así, existen tres diferentes paquetes estadísticos construidos sobre Lisp-Stat con diferentes capacidades y modo de uso. Estos son ViSta (Young and Bann, 1997), R<sup>8</sup> (Weisberg, 1997) y Axis (Stine, 1997). Todos ellos comparten el que, al subyacer Lisp-Stat a todos ellos, es posible ampliarlos de una manera relativamente fácil.

Existen implementaciones de Lisp-Stat para UNIX, Microsoft Windows y Mac OS. Es gratuito y puede ser obtenido a través de Internet, junto con información acerca de su

---

<sup>8</sup> No confundir con el anterior R.

uso. Existe código para hacer diversos análisis estadísticos también disponible en diversos lugares de la red Internet.

Puesto que ambos lenguajes son relativamente similares es posible dar una descripción de las características que poseen en común en cuanto a funcionamiento. Estas son:

a) *Interactividad*: Ambos sistemas poseen un interpretador que puede ser utilizado como una calculadora estadística. Por ejemplo una expresión del tipo (media x) en donde x es una variable produciría que se mostrara la media.

b) *Aritmética vectorizada*: Generalmente los lenguajes de programación incorporan una gran cantidad de funciones matemáticas útiles para realizar análisis estadísticos. Por ejemplo, tenemos logaritmos, exponentes, etc. Normalmente, estas funciones están concebidas para ser aplicadas a un número cada vez. Los paquetes estadísticos en cambio suelen aplicar esas funciones a todos los miembros de un vector. Esto se denomina aritmética vectorizada. Lisp-Stat y S-Plus toman este camino. Por ejemplo, en Lisp-Stat (+ x y) en donde x e y son dos vectores producirá el sumatorio de cada par de valores de x e y.

c) *Programación funcional*: Las expresiones tecleadas interactivamente pueden también ser agrupadas y ejecutadas en un bloque. Este bloque puede recibir un nombre y convertirse en una nueva función del programa que no difiere en su utilización de las proporcionadas originalmente por el entorno de programación. En realidad, programar en Lisp-Stat y S-Plus implica básicamente construir nuevas funciones más complejas que llaman a otras más básicas. Al final, el número de funciones disponibles es enorme. Un ejemplo podría ser el siguiente. Si tenemos que (media x) nos produce la media de una variable, podemos hacer (- x (media x)). El resultado obtenido es restar a cada valor de x la media de la variable, obteniendo las puntuaciones diferenciales (esto se consigue por que la definición de la operación resta en Lisp-Stat es vectorial, lo cual implica que se aplica a cada elemento de una lista o vector). Esta expresión podría convertirse en una nueva función mediante:

```
(defun punt-dif (x) (- x (media x)))
```

Después de esta definición, la expresión (punt-dif x) produce como resultado automático las puntuaciones diferenciales de x.

Un objetivo de la programación funcional es evitar al máximo la asignación de valores a variables como método para guardar resultados intermedios. Ello supuestamente hace al código producido más fácil de entender y revisar (Tierney, 1990).

d) *Estructuras de datos flexibles*. Esto ya ha sido comentado en otros lugares. Tanto en S-Plus como en Lisp-Stat existen tres estructuras que no están disponibles habitualmente para el usuario en paquetes estadísticos y que constituyen ayudas muy interesantes para la programación. Estas son matrices multidimensionales (con dos o más dimensiones), listas y objetos. No poder acceder a estas estructuras limita, o al menos hace más complejas, la capacidades de programación en paquetes estadísticos.

e) *Resultados como datos frente a presentación en pantalla*. S-Plus y Lisp-Stat tienden a producir los resultados en forma de variables que posteriormente pueden ser accedidas y manipuladas de nuevo. Por ejemplo, si se solicita los estadísticos descriptivos de unos datos multivariados se puede obtener un vector con las medias, otro con las desviaciones típicas, etc. Esos vectores pueden ser consultados o pueden ser, por ejemplo, representados gráficamente (por ejemplo un diagrama de medias v. desviaciones típicas en datos con supuestamente escala similar podría ser interesante). Por contra los paquetes estadísticos tenderán a producir un resultado en pantalla, pero lo obtenido no puede ser reaprovechado para nuevos cálculos.

f) *Posibilidad de incrementar el sistema y apoyo a la programación experimental*. Ambos sistemas ofrecen la posibilidad de incorporar las nuevas funciones que se van construyendo al conjunto de funciones disponibles. Esto significa que si cierto procedimiento estadístico no está disponible en un momento dado resulta fácil incorporarlo al resto si uno decide construirlo por sí mismo.

Otra característica muy importante es que estos sistema no fuerzan a un estilo de programación rígido, en el que es necesario seguir una determinada estructura si se quiere garantizar el funcionamiento del código. Por el contrario, estos lenguajes llevan a la construcción de una serie de pilares básicos que posteriormente van combinándose hasta formar la estructura final. Esto ha venido a denominarse estructura arriba-abajo v. estructura abajo-arriba. En general, la estructura arriba-abajo se entiende que produce mejores resultados cuando los problemas están bien entendidos y lo que se desea es optimizar el resultado final. Por el contrario, en problemas todavía mal definidos, una estructura abajo arriba permite un desarrollo más equilibrado en el que cada parte es atacada y solucionada por separado hasta lograr zanjarlos.

g) *Gráficos*. En los últimos tiempos se han propuesto una gran variedad de métodos de análisis gráfico de datos (Chambers, et al., 1983). Este tipo de análisis ha recibido un

gran apoyo en el hecho que las capacidades gráficas de los computadores han experimentado grandes mejoras. Muchos de los paquetes estadísticos desarrollados en los años 60 carecen de esas capacidades y, aunque sus nuevas versiones ofrecen mejoras en esa dirección, es muy posible que carezcan de una base conceptual que les limita en cuanto a la potencialidad a alcanzar. Tanto Lisp-Stat como S-Plus ofrecen capacidades gráficas muy interesantes, las cuales tienen un alto grado de flexibilidad, de tal modo que resulta posible para el usuario concebir y desarrollar nuevos tipos de gráficos.

h) *Orientación a objetos*. La orientación a objetos ha supuesto una gran revolución dentro del mundo de la programación que en los últimos años ha estado alcanzando a los desarrollos estadísticos (Stine and Fox, 1997a). Gracias a ella, los usuarios pueden aprovechar objetos previamente desarrollados y añadirles nuevas funciones (métodos en la terminología de la orientación a objetos) que los adapten a las necesidades del momento. Este paradigma es especialmente natural para la programación en sistemas basados en interfaces de usuario gráficos los cuales se están convirtiendo en los últimos años en los más comunes así como apoyo para la programación experimental e incremental.

## 8.4. Conceptos básicos de programación

Señalar cuales son los aspectos básicos de la programación es difícil puesto que cada lenguaje presenta conceptos diferentes que llevan a que básico en uno no lo sea en otro. En nuestro caso hemos optado por tomar como modelos los entornos de programación anteriormente descritos (Lisp-Stat y S-Plus) puesto que ambos están más centrados en el tipo de tareas que consideramos en este texto (las estadísticas).

Los temas que expondremos son los siguientes: a) Asignamiento de valores a variables, b) Predicados, c) Evaluación condicional, d) Bucles y evaluación caso a caso (mapping), e) Input/Output, f) Estructuras de control, g) Tipos de datos y h) Programación orientada a objetos.

a) Asignamiento de valores a variables. El asignamiento nos permite utilizar un símbolo en referencia a unos valores. Por ejemplo, en Lisp-Stat haríamos.

```
(def x (list 1 2 3))
```

Después de esto podemos escribir.

```
(mean x)
```

Y el sistema automáticamente contestará:

```
>2
```

Asignar unos datos normalmente también implica definir el tipo de datos al que nos referimos. Antes hemos asignado a `x` una lista (comando `list`) pero también podríamos especificar:

```
(def x (vector 1 2 3))
```

En este caso `x` es un vector. Generalmente cuando el número de valores a usar es pequeño no existen muchas diferencias entre uno u otro tipo de datos. Cuando este aumenta las operaciones sobre vectores ejecutan mucho más rápido. No obstante, las listas son mucho más flexibles y admiten manipulaciones más sofisticadas.

b) Predicados y valores lógicos. Un valor lógico es el que resulta de determinar si una condición es verdadera o falsa. Los predicados son funciones que producen un valor lógico. Para comparar números están disponibles generalmente los siguientes predicados:

```
= igual que
```

```
< menor que
```

```
> mayor que
```

```
<= menor o igual que
```

```
>= mayor o igual que
```

Por ejemplo, en Lisp-Stat escribiríamos:

```
(< 2 3)
```

Con el resultado de:

```
T (verdadero)
```

Los operadores lógicos permiten la construcción de predicados más complejos. Generalmente existen tres operadores:

- AND ( $\gamma$ ). p.e. (AND (> 3 2) (< 2 1)) -> FALSO
- OR ( $\cup$ ). p.e. (OR (> 3 2) (< 2 1)) -> VERDADERO
- NOT (no) p.e. (NOT (OR (> 3 2) (< 2 1))) -> FALSO

c) Evaluación condicional: La evaluación condicional nos permite emprender acciones en función de comparaciones lógicas. La estructura de evaluación condicional más importantes es normalmente:

```
SI <predicado> ENTONCES <acción> SINO <acción 2>
```

Por ejemplo:

```
(SI (< (media x) 10) (media x) (print "hay un error"))
```

El comando anterior indica que si la media de una variable es menor de 10 entonces muestra la media de la variable. En caso contrario imprime en pantalla "hay un error".

d) Bucles y evaluación individualizada. A menudo se dice que la capacidad de los ordenadores está fundamentalmente en hacer muy rápidamente cosas simples. Los bucles son precisamente estructuras dirigidas a la repetición de esas acciones simples. La forma que tiene un bucle es:

```
Repetir <acción> hasta <predicado>
```

Además, los bucles generalmente incorporan un *contador*, el cual se incrementa por cada repetición. Por ejemplo, supongamos que queremos tener una variable con los números enteros desde el 1 al 100. En pseudocódigo escribiríamos:

```
REPETIR DESDE x=1 HASTA 100
  AÑADIR lista_números x
```

VOLVER (SI X<=100)

En este caso x es el contador. El programa empieza el bucle en REPETIR y asigna a x el valor de 1. A continuación añade el valor de x a la lista de números y, puesto que x es menor que 100, volvería a empezar pero asignando a x en este caso el valor de 2. Esto se repetiría hasta que x alcanzara el valor de 100, momento en que el programa terminaría con el bucle.

A veces los bucles están anidados, de modo que tenemos uno interior y otro exterior por así decir. Por ejemplo, un problema muy habitual es cambiar todos los elementos de una matriz. Supongamos que tenemos la matriz siguiente:

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix}$$

Y queremos sumar uno a todos los valores para obtener

$$\begin{vmatrix} 2 & 3 \\ 4 & 5 \end{vmatrix}$$

Utilizando bucles esto podría hacerse del siguiente modo:

```
REPETIR DESDE x=1 HASTA 2  
  
    REPETIR DESDE y=1 HASTA 2  
  
        Sumar MATRIZ [x y] 1  
  
    VOLVER (SI y<=2)  
  
VOLVER (SI x<=2)
```

Este código iría elemento por elemento de la matriz sumando 1 a todos ellos.

Los bucles son una estructura de programación muy interesante, aunque su abuso produce un código muy difícil de entender. Esto se agrava cuando se anidan varios bucles y se llevan a cabo asignaciones a variables. En esos casos no es fácil determinar cual es el valor de las variables en cada momento y los errores son muy difíciles de determinar.

Por otro lado, en lenguajes interpretados en los que cada instrucción es evaluada individualmente el código puede resultar muy poco eficiente. Por ello, tanto S-Plus como Lisp-Stat incorporan funciones que, al ser aplicadas a variables con varios elementos,



trabajan para cada uno de ellos. Por ejemplo, si la matriz anterior tuviera asignado el símbolo *M*, en Lisp-Stat podemos hacer:

```
(+ 1 M)
```

El resultado es sumar uno a cada elemento de la matriz tal y como hemos hecho antes mediante los dos bucles.

En ocasiones, la función que estamos aplicando no está diseñada para trabajar con varios elementos simultáneamente. Supongamos que queremos calcular las medias de las columnas de una matriz. En Lisp-Stat podría escribirse:

```
(map-elements #'mean (column-list M))
```

Column-list extrae las columnas de una matriz siendo cada una de ellas un "elemento". El comando map-elements aplica la media (mean) a cada uno de los elementos, obteniéndose un vector con las medias de las columnas.

#### e) Input/output

Una de las partes esenciales de nuestros programas es mostrar y conservar los resultados. Desde el punto de vista del computador, ambas operaciones son bastante semejantes y equivalen a una noción general de imprimir, bien sobre la pantalla o bien sobre un sistema de almacenamiento como puede ser un disco duro. Ambas operaciones presentan elementos comunes, aunque también tienen peculiaridades propias.

La primera noción sobre la impresión en pantalla es que no todos los pasos intermedios aparecerán en pantalla. De hecho, cuando utilicemos funciones siempre es necesario tener en cuenta el valor que producen finalmente. Por ejemplo, si definimos la siguiente función en Lisp-Stat:

```
(defun descriptives (m)
  (mean m)
  (max m)
  (min m)
)
```

Y usamos,

```
(descriptives (list 1 2 3))
```

El resultado será 1. Es decir, la última de las operaciones de la función `descriptives`. Si quisiéramos obtener los tres resultados tendríamos que redefinir la función `descriptives` para que aparezca del siguiente modo:

```
(defun descriptives (m)
  (list
    (mean m) (max m) (min m)
  ))
```

El resultado será ahora:

```
(2 3 1)
```

De todos modos, el resultado de las funciones no está hecho para los usuarios sino más bien para uso interno del programa. Una función más apropiada es, en Lisp-Stat, `format`. Por ejemplo,

```
(format t "Tabla pequeña ~%~10.3g ~10.3g %~10.3g ~10.3g" 1.2
        3.1 4.7 5.3)
```

Da como resultado:

```
Tabla pequeña
1.2 3.1
4.7 5.3
```

En breve, la parte entrecomillada instruye al programa la forma en que se debe imprimir la segunda parte. El texto se imprime como está y el símbolo `%` obliga a un cambio de línea. El símbolo `~g` instruye al programa a usar la notación decimal o científica que ocupe menos espacio para ese número. `10` es el número de espacios entre columnas de datos y `3` el número de decimales que utiliza.

Este tipo de formato está especialmente diseñado para datos numéricos y es el que permite crear columnas de datos con la misma presentación y aspecto de manera consistente.

En cuanto a los archivos, éstos permiten guardar la información entre sesiones, así como utilizar archivos de gran tamaño. En general, para utilizar estas instrucciones es necesario tener en cuenta dos aspectos, el camino que señala a un archivo ("path") y la dirección de la información. La forma de nombrar los caminos depende del sistema operativo y no del lenguaje de programación generalmente, así que, por ejemplo, lo siguiente puede ser un camino válido en el sistema operativo de Microsoft (DOS o Windows):

```
c:/programación/ejemplo.lsp
```

Otros sistemas operativos pueden utilizar otros métodos.

Una cuestión que es necesario manejar cuando se programa es si el archivo sobre el que se está trabajando es de lectura o de escritura. Un archivo de lectura sólo puede ser utilizado para extraer información y uno de escritura para volcar información. Esto impide que se produzcan conflictos entre operaciones opuestas.

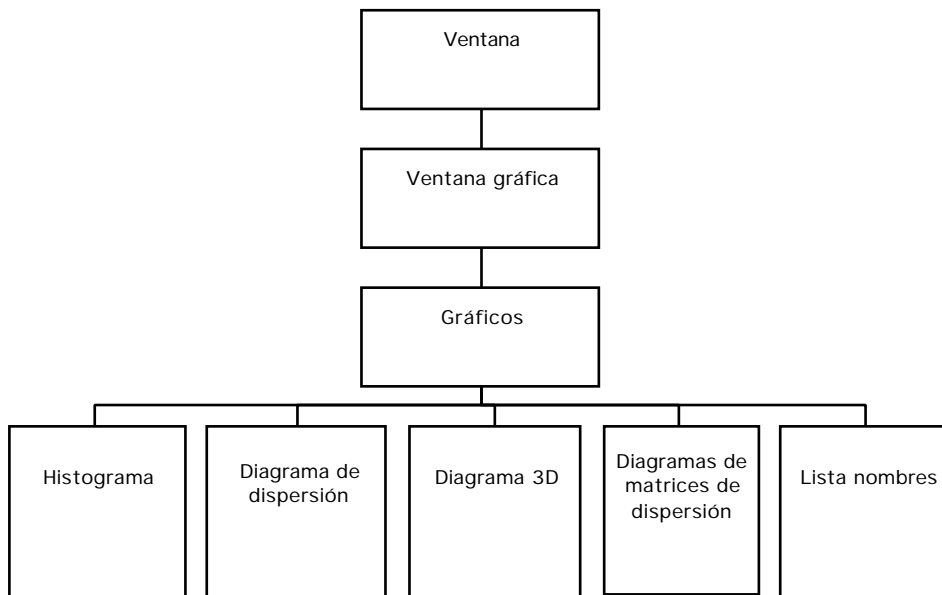
#### f) Programación orientada al objeto

En los últimos tiempos la programación ha estado dominada por el paradigma de orientación a objetos. Entre sus ventajas podemos citar que gracias a él se favorece mucho la reusabilidad de código y que, además, resulta natural para la programación de entornos de tipo gráfico.

Un ejemplo ayudará a entender su importancia. Supongamos que tenemos código que nos permite hacer un gráfico de un conjunto de datos. Inicialmente tenemos tres tipos de datos (univariados, bivariados y multivariados) y escribimos código para que realice un gráfico apropiado (histograma, diagrama de dispersión, gráfico 3D y matrices de diagramas de dispersión). Supongamos que luego queremos añadir un gráfico para un cuarto tipo de datos: una serie temporal. Puesto que el código anterior ya está construido sería necesario modificarlo para adaptarlo a este nuevo gráfico. Ello nos obliga a entender un código que, posiblemente, quizás no fue escrito por nosotros mismos corriendo el riesgo de hacer que deje de funcionar. Una alternativa es que el código sea capaz de entender un mensaje indicando el tipo de datos que le estamos enviando y a partir de él decida que rumbo debe de emprender. Así añadir este nuevo gráfico consiste en hacer que el programa entienda un nuevo mensaje y sepa como responder a él, no siendo necesario modificar el código anterior.

Los conceptos de métodos y de envío de mensajes son fundamentales en la programación orientada al objeto. Un tercer concepto de importancia es el de herencia. Un nuevo objeto puede tomar prestado de otros objetos existentes todos o varios de sus métodos. A estos métodos se le pueden añadir otros nuevos sin por ello modificar los ya existentes que funcionan correctamente. Nuestros objetos consistirán normalmente de los métodos de otros objetos previos más los que nosotros añadamos.

Esto nos lleva a un cuarto concepto: la jerarquía de objetos. Puesto que los métodos más especializados van tomando métodos de los más generales resulta posible ponerlos en orden con respecto a una jerarquía de abstracción-particularidad. Por ejemplo, la jerarquía de objetos gráficos en Lisp-Stat es la siguiente:



En este ejemplo, una ventana tiene una conducta y unas acciones propias las cuales son heredadas por las ventanas de gráficos. Las ventanas de gráficos incorporan métodos para dibujar líneas, puntos, círculos, etc. Los gráficos estadísticos añaden métodos para dibujar gráficos a partir de datos y para manipular esos gráficos. Por último, los diversos objetos se especializan en datos univariados, bivariados, trivariados y multivariados. Un último objeto permite trabajar con listas de nombres (por ejemplo para indicar las etiquetas de observaciones o de variables).

# Referencias

- Autor (1983). *Webster's encyclopedic unabridged dictionary of the english language*. Gramercy books.
- Afifi, A. A. y Clark, V. (1984). *Computer-Aided Multivariate Analysis*. New York: Van Nostrand Reinhold Company.
- Aldenderfer, M. S. y Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills: Sage Publications.
- Aldrich, J. (1984). *Linear probability, logit, and probit models*. Sage.
- Amon, J. (1991). *Introducción al análisis multivariante (cálculomatricial)*. Barcelona: PPU.
- Arce, C. (1994). *Técnicas de Construcción de Escalas Psicológicas*. Madrid: Síntesis.
- Arnau, J. (1998). Metodología de la Investigación y Diseño. En Arnau y Carpintero (Eds.), *Tratado de Psicología General I, Historia, Teoría y Método*. Madrid: Alhambra.
- Atkinson, A. C. y Cox, D. R. (1982). Transformations. En Kotz y Johnson (Eds.), *Encyclopedia of Statistical Sciences*. New York: Wiley y Sons.

- Bajo Delgado, M. T. y Cañas Molina, J. J. (1991). *Ciencia Cognitiva*. Madrid: Editorial Debate.
- Bakeman, R. y Quera, V. (1995). *Analyzing Interaction sequential analysis with SDIS and GSEQ*. Cambridge: Cambridge University Press.
- Barnett, V. y Lewis, T. (1994). *Outliers in Statistical Data*. Chichester: John Wiley and Sons.
- Barton, C., Hatcher, C., Schurig, K., Marciano, P., Wilcox, K. y Brooks, L. (1991). Managing data entry of a large-scale interview project with optical scanning hardware and software. *Behaviour Research Methods, Instruments and Computers*, **23**(2), 214-218.
- Beaton, A. E. (1964). The use of special matrix operations in statistical calculus. *Educational Testing Service Research Bulletin*, **RB-64-51**.
- Becker, R. A. y Chambers, J. M. (1984). *S: An interactive Environment for Data Analysis and Graphics*. Belmont, CA: Wadsworth.
- Becker, R. A., Chambers, J. M. y Wilks, A. R. (1988). *The new S language: A programming environment for Data Analysis and Graphics*. Pacific Grove, CA: Wadsworth.
- Bergman, L. y Magnusson, D. (1990). General issues about data quality in longitudinal research. En D. y Bergman Magnusson L. R. (Eds.), *Data quality in longitudinal research* (pp. 1-31). Cambridge: Cambridge University Press.
- Boulton, D. y Hammersley, M. (1996). Analysis of unstructured data. En R. Sapsford and Jupp V. (Eds.), *Data Collection and Analysis* London: Sage.
- Bourque, L. B. y Clark, V. A. (1992). *Processing Data: The Survey Example*. Sage.
- Bozdogan, H. y Ramirez, D. E. (1988). UTRANS and MTRANS: Marginal and Joint Box-Cox Transformation of Multivariate Data to "Near" Normality. *Multivariate Behavioral Research*, **23**, 131-132.
- Bryk, A., y Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis*. Newbury Park, CA: Sage.
- Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, **B22**(302-306).
-

- Carroll, J. M. & Thomas, J. C. (1982). Metaphor and the cognitive representation of computing systems. *IEEE Transactions on Systems, Man & Cybernetics*, (2), 107-115.
- Carroll, J. D. y Arabie, P. (1980). Multidimensional Scaling. *Annual Review of Psychology*, **31**, 607-649.
- Chambers, J. M., Cleveland, W. S., Kleiner, B. y Tukey, P. A. (1983). *Graphical methods for data analysis*. Pacific Grove, California: Wadsworth y Brooks.
- Cleveland, W. S. y McGill, M. E. (1988). *Dynamic Graphics for Statistics*. Belmont: Wadsworth y Brooks.
- Cochran, W. G. (1963). *Sampling Techniques*. New York: Wiley.
- Coombs, C. H. (1964). *A theory of data*. New York: Wiley.
- Cowles, M. (1989). *Statistics in Psychology: An historical perspective*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Data Analysis Products Division (1997). *S-Plus Programmer's Guide*. Seattle, WA:
- Date, C. J. (1993). *Sistemas de bases de datos*. Wilmington: Addison-Wesley Iberoamericana.
- Davidson, F. (1996). *Principles of statistical data handling*. Thousand Oaks: Sage Publications.
- Dixon, W. J. (1992). *BMDP Statistical Software Manual*. Los Angeles: University Press of California.
- Dixon, W. J. (Ed.). (1981). *BMDP Statistical Software*. Berkeley: University of California Press.
- Eliason, S. R. (1993). *Maximum Likelihood: Estimation and Practice*. Newbury Park, CA: Sage.
- Emerson, J. D. y Stoto, M. A. (1983). Transforming data. En D. C. Hoaglin, Mosteller, F. y Tukey, J. W. (Eds.), *Understanding Robust and Exploratory Data Analysis*. New York: Wiley and Sons.
-

- Emerson, J. D. (1991). Introduction to transformations. En D. C. Hoaglin, Mosteller, F. y Tukey, J. W. (Eds.), *Fundamentals of exploratory analysis of variance* New York: Wiley y Sons.
- Fellegi, I. P. y Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of The American Statistical Association*, **71**(353), 17-35.
- Freedland, K. E. y Carney, R. M. (1992). Data management and accountability in behavioral and biomedical research. *American Psychologist*, **47**(5), 640-645.
- Gabriel, K. R. (1986). Biplot display of multivariate matrices for inspection of data and diagnosis. En V. Barnett (Eds.), *Interpreting multivariate data*, Chichester (UK): Wiley.
- Gharamani, Z. y Jordan, M. I. (1994). *Learning from incomplete data* (Report Number 1509). Massachusetts Institute of Technology.
- Gil, J., García, E. y Rodríguez, G. (1998). Análisis de datos cualitativos. En J. Renom Pinsach (Eds.), *Tratamiento informatizado de datos* (pp. 41-67). Barcelona: Masson.
- Goldstein, H. y McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, **53**(4), 455-467.
- Graham, J. W. (1998). Intercambio por correo electrónico.
- Graham, J. W., Hofer, S. M. y MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, **31**(2), 197-218.
- Green, R. F. (1982). Outlier-Prone Distribution. En Kotz y Johnson (Eds.), *Encyclopedia of Statistical Sciences*. New York: Wiley and Sons.
- Green, W. H. (1993). *Econometric Analysis*. New York: MacMillan.
- Harrison, P. R. (1990). *Common Lisp and Artificial Intelligence*. Englewood Cliffs, NJ: Prentice Hall.
- Hawkins, D. M. (1982). Outliers. En Kotz y Johnson (Eds.), *Encyclopedia of Statistical Sciences*. New York: Wiley y Sons.
- Hoaglin, D. C., Mosteller, F. y Tukey, J. W. (1991). *Fundamentals of Exploratory Analysis of Variance*. New York: Wiley y Sons.
-



- Hoaglin, D. C., Mosteller, F. y Tukey, J. W. (Ed.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- Howes, A. (1994). A model of the acquisition of menu knowledge by exploration. In B. Adelson, S. Dumais and J. Olson (Ed.), *CHI'94*, (pp. 445-451). Boston, Massachusetts: ACM.
- ICPSR. (1997). *ICPSR Guide to social science data preparation and archiving*. Interuniversity Consortium for Political and Social Research.
- Jaffe, J. A. (1994). *Mastering the SAS system*. New York: Van Nostrand Reinhold.
- Klinke, S. (1997). *Data structures for computational statistics*. Berlin: Physica-Verlag.
- Kreft, I. E. (1995). Special issue: Hierarchical Linear Models: Problems and Prospects. *Journal of Educational and Behavioral Statistics*, **20**(2), 109-240.
- Krosnick, J. A. y Fabrigar, L. R. (En prensa). Open and closed questions. En J. A. Krosnick and L. R. Fabrigar (Eds.), *Designing questionnaires to measure attitudes (tentative title)* (pp. 1-45). New York: Oxford University Press.
- Krosnick, J. A. y Fabrigar, L. R. (En prensa). *Designing good questionnaires: Insights from Cognitive and Social Psychology*. New York: Oxford University Press.
- Little, R. J. A. y Rubin, D. A. (1987). *Statistical analysis with missing data*. New York: Wiley y Sons.
- López, J. J., Ato, M., Rabadán, R. y Galindo, F. (1997). Sistemas de codificación y regresión logística. In *V Congreso de Metodología de las Ciencias Sociales y Humanas*. Sevilla:
- Losilla Vidal, J. M., Navarro Pastor, J. B. y Vives Brosa, J. (1997). *Diseño de bases de datos relacionales. MS-Access 97*. Barcelona.
- Mattox, J. R., O. Dwyer, W. y Leeming, F. C. (1997). Making traffic crash data useful through rapid record entry and analysis. *Journal of Safety Research*, **28**(4), 221-231.
- Miles, M. B. y Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks: Sage.
- Naus, J. I. (1982). Editing Statistical Data. En Kotz y Johnson (Eds.), *Enciclopedia of Statistical Sciences*. New York: Wiley y Sons.
-

- Naus, J. I., Johnson, T. G. y Montalvo, R. (1972). A probabilistic model for Identifying Errors in Data Editing. *Journal of the American Statistical Association*, **67**(340), 943-950.
- Norman, D. A. (1987). Cognitive engineering-Cognitive Science. En J. M. Carroll (Eds.), *Interfacing Thought: Cognitive aspects of Human-Computer Interaction* (pp. 325-336). Cambridge: A Bradford Book.
- Norusis, M. (1988). *SPSS-X Tables*. Chicago: SPSS Inc.
- Norusis, M. (1990a). *SPSS Advanced Statistics. User's Guide*. Chicago: SPSS Inc.
- Norusis, M. (1990b). *The SPSS guide to data analysis*. Chicago: SPSS Inc.
- Norusis, M. (1990c). *SPSS Base System*. Chicago: SPSS Inc.
- Norusis, M. (1993). *SPSS for Windows: Base System User's Guide Release 6.0*. Chicago: SPSS Inc.
- Orchard, T. y Woodbury, M. A. (1972). A missing information principle: theory and applications. In *Proceedings of the 6th Berkely Symposium on Mathematical Statistics and Probability*, 1 (pp. 697-715).
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research..* New York: Holt, Rinehart y Winston.
- Perea, M. y Pitarque, A. (1990). Simulación en Psicología. En S. Algarabel y J. Sanmartín (Eds.) *Métodos Informáticos aplicados a la Psicología*, Valencia: Pirámide.
- Pylyshyn, Z. W. (1988). *Computación y Conocimiento*. Madrid: Debate.
- Quiñones Vidal, E., Garcia Sevilla, J. y Pedraja Linares, M. J. (1989). El uso de instrumentos en la investigación psicológica. En J. J. Mayor and J. L. Pinillos (Eds.), *Tratado de Psicología General* (pp. 373-390). Madrid: Alhambra.
- Raymond, M. R. y Roberts, D. M. (1983). Development and validation of a foreign language attitude scale. *Educational and Psychological Measurement*, **43**, 1239-1246.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley and Sons.
- Rubin, D. B. (1991). EM and Beyond. *Psychometrika*, **56**(2), 241-254.
-

- Sanderson, P. y Fisher, C. (1994). Exploratory Sequential Data Analysis: Foundations. *Human-Computer Interaction*, **9**(3-4), 251-318.
- Sanderson, P., Scott, J., Johnston, T., Mainzar, J., Watanabe, L. y James, J. (1994). MacSHAPA and the Enterprise of Exploratory Sequential Data Analysis. *International Journal of Human-Computer Studies*, **41**, 633-681.
- Sanmartín, J. y Algarabel, S. (1990). Introducción. En S. Algarabel and J. Sanmartín (Eds.), *Métodos informáticos aplicados a la Psicología* (pp. 19-36). Madrid: Pirámide.
- Sanmartín, J., Meliá, J. L. y Soler, M. J. (1990). Estructura y funcionamiento de un ordenador. En S. y Sanmartín Algarabel J. (Eds.), *Métodos informáticos aplicados a la Psicología* Valencia: Pirámide.
- Saris, W. E. y Gallhofer, N. I. (1998). *Formulation and classification of questions*. Department of Methods and Techniques for Social Science Research. Netherlands Organization for Scientific Research.
- Saris, W. E. (1991). *Computer-Assisted Interviewing*. Newbury Park: Sage.
- Sarle, W. S. (1995). Measurement theory: Frequently asked questions. En *Disseminations of the International Statistical Applications Institute, 4th edition* (pp. 61-66). Wichita: ACG Press.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman y Hall.
- Smith, S. L. y Mosier, J. N. (1986). *Guidelines for designing user interface software*. Bedford Massachusetts: Approved for public release.
- Steele, G. L. (1984). *Common Lisp. The Language*. Digital Press.
- Stine, R. y Fox, J. (1997a). Editor's Introduction. En R. Stine y J. Fox (Eds.), *Statistical Computing Environments for Social Research* (pp. 1-17). Thousand Oaks: Sage.
- Stine, R. y Fox, J. (1997b). *Statistical computing environments for social research*. Thousand Oaks: Sage.
- Stine, R. (1997). Axis: An Extensible Graphical User Interface. En R. y Fox Stine J. (Eds.), *Statistical Computing Environments for Social Research* (pp. 175-192). Thousand Oaks: Sage.
-

- Swift, B. (1996). Preparing Numerical Data. En R. Sapsford and V. Jupp (Eds.), *Data Collection and Analysis* (pp. 153-183). London: Sage Publications.
- Tabachnik, B. G. y Fidell, L. S. (1989). *Using Multivariate Statistics*. Northridge: HarperCollins.
- Tierney, L. (1989). *XLISP-STAT: A Statistical Environment based on the XLISP language* (528). University of Minnesota. School of Statistics.
- Tierney, L. (1990). *LISP-STAT: An object-oriented environment for statistical computing and dynamic graphs*. New York: Wiley.
- Trawinski, J. M. y Bargmann, R. W. (1964). Maximum likelihood with incomplete multivariate data. *Ann. Math. Statist.*, **35**, 647-657.
- Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Massachusetts: Addison Wesley.
- Valero, P., Molina, G. & Sanmartín, J. (1992). El uso de ordenadores en la enseñanza de la Psicología: Selección del material, infraestructura e integración en la docencia. In *II Congreso Iberoamericano de Psicología*. Madrid.
- Valero, P., Molina, G. & Sanmartín, J. (1992). El uso de ordenadores en la enseñanza de la Estadística en Psicología. In *II Congreso Iberoamericano de Psicología*. Madrid.
- Velleman, P. F. y Wilkinson, L. (1993). Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician*, **47**(1), 65-72.
- Velleman, P. F. (1995). *Data Desk Handbook*. Ithaca: Data Description Inc.
- Walker, N. y C., R. (1993). Aggregation bias and the use of regression in evaluating models of human performance. *Human Factors*, **35**(3), 397-411.
- Weber, R. P. (1985). *Basic content analysis*. Beverly Hills: Sage.
- Weisberg, S. (1997). The R-Code: A graphical paradigm for regression analysis. En R. y Fox Stine J. (Eds.), *Statistical Computing Environments for Social Research* (pp. 193-206). Thousand Oaks: Sage.
-

- Weitzman, E. A. y Miles, M. B. (1995). *Computer programs for qualitative data analysis: A software sourcebook*. Thousand Oaks: Sage.
- West, M. y Winkler, R. L. (1991). Data Base Error Trapping and Prediction. *Journal of the American Statistical Association*, **86**(416), 987-996.
- White Paper. (1997). *Missing data: the hidden problem*. SPSS, Inc.
- Wilkinson (1989). *SYSTAT: The System for Statistics*. Evanston, IL: Systat, Inc.
- Willitts, J. (1992). *Database design and construction*. London: Library Association Publishing.
- Young, F. y Bann, C. M. (1997). ViSta :A Visual Statistics System. En R. y Fox Stine J. (Eds.), *Statistical Computing Environments for Social Research* (pp. 207-235). Thousand Oaks: Sage.
- Young, F. W. y Hamer, R. M. (1987). *Multidimensional scaling: History, Theory and Applications*. Hillsdale, New Jersey: LEA.
- Young, F. W. y Harris, D. F. (1990). Multidimensional scaling: procedure alsca. En Marija Norusis (Eds.), *SPSS Base System User's Guide* (pp. 396-461). Chicago :Illinois: SPSS Inc.
- Young, F. W. (1996). *ViSta: Developing Statistical Objects. Research Memorandum 96-1.*. Chapel Hill, NC: L.L. Thurstone Psychometric Laboratory, University of North Carolina.
-