



UNIVERSIDAD NACIONAL
AUTÓNOMA

11214
200
UNIDAD ACADÉMICA DE LOS CICLOS
PROFESIONAL Y DE POSGRADO DEL C.C.H.
INSTITUTO DE INVESTIGACIONES EN
MATEMÁTICAS APLICADAS Y SISTEMAS.

MANEJO DE ARCHIVOS DE DATOS CON EL
MÓDULO "DATA MANAGER" DEL PAQUETE
ESTADÍSTICO BMDP

Tesina que para obtener el Diploma de la
Especialización en Estadística Aplicada
presenta la:

ACT. ROCÍO HAYDEÉ CAMERO MARTÍNEZ

Asesorada por:

M. en C. I. RAFAEL MADRID RÍOS.

MÉXICO D.F.

1993.

**TESIS CON
FALLA DE ORIGEN**



UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

DERECHOS RESERVADOS © PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Í N D I C E

PRÓLOGO.....1

CAPÍTULO 1. INTRODUCCIÓN.

1.1 Justificación del Trabajo.....2

1.2 Ejemplos.....3

1.2.1 Investigación 1. Relación entre Migración y
Salud Mental.....3

1.2.2 Investigación 2. Estimación para Diseños de
Muestreo.....4

1.2.3 Investigación 3. Efecto de trasplante de teji-
do cerebral embrionario en la recuperación fun-
cional de ratas con Hemiplejía Experimental...5

1.3 Objetivos del Trabajo.....7

CAPÍTULO 2. IDEAS GENERALES PARA EL MANEJO DE ARCHIVOS CON EL MÓDULO DATA MANAGER (DM) DEL PAQUETE ESTADÍSTICO BMDP.

2.1 Conceptos afines con el módulo DM del BMDP.....9

2.2 Estructura de un programa de instrucciones de
lenguaje BMDP, utilizando el módulo DM.....16

2.3 Descripción del Módulo Data Manager.....18

CAPÍTULO 3. APLICACIÓN DEL DATA MANAGER.

3.1 Investigación 1.....53

Problema por resolver:

3.1.1 Creación y mezcla de archivos.....54

3.1.2 Mapa de datos faltantes.....56

3.2 Investigación 2.....60

Problema por resolver: Estimación para

3.2.1 Muestreo estratificado con selección aleatoria simple sin reemplazo.....61

3.2.2 Muestreo bietápico con selección con probabilidad proporcional al tamaño con reemplazo en la primera etapa y con muestreo aleatorio simple sin reemplazo en la segunda etapa.....71

3.3 Investigación 3.....79

Problema por resolver:

3.3.1 Transposición de datos.....84

CAPÍTULO 4. CONCLUSIONES.....98

REFERENCIAS Y BIBLIOGRAFÍA103

PRÓLOGO

En la aplicación de la Estadística con frecuencia existen problemas relacionados con el manejo de datos y con la preparación de estos para la aplicación de los métodos de análisis. Para ello se forman archivos de datos, que pueden ser demasiado grandes debido a un gran número de muestras (registros-observaciones-casos) o a un gran número de variables medidas en cada muestra.

Es precisamente en este contexto en que el uso del módulo DATA MANAGER (DM) del paquete estadístico BMDP, es un gran auxiliar en la aplicación de la estadística, ya que DM es un sistema potente de manipulación de datos que permite 'poner a punto' la información, antes de usar los procesos estadísticos requeridos.

Así, una vez organizada la información con el DM, se encontrará lista para ser procesada por cualquiera de los demás módulos del BMDP, los cuales son compatibles con el módulo DM.

El presente trabajo está estructurado de la manera siguiente:

En el capítulo 1 se encuentra la Introducción, donde se justifica la necesidad del manejo de archivos en la aplicación de la Estadística. Para ello se describen problemas por resolver en tres investigaciones y se definen los objetivos que se pretenden alcanzar con la escritura del presente trabajo.

En el capítulo 2 se explica en qué consiste la preparación ('puesta a punto') de la información con el módulo DM, tomando en cuenta los elementos básicos de un programa de instrucciones de lenguaje de BMDP. En particular se explican los párrafos del DM y sus respectivas instrucciones. Se definen conceptos que se mencionan en el trabajo.

En el capítulo 3 se ejemplifica el potencial del DM en tres investigaciones que en la práctica se han analizado y que han sido referidas en el capítulo 1.

En el capítulo 4 se presentan las conclusiones, donde además se plantean otras situaciones de interés donde el DM puede ser utilizado.

CAPÍTULO 1.

INTRODUCCIÓN

1.1 Justificación del Trabajo.

En la aplicación de la estadística es necesario contar con una serie de herramientas que permitan un análisis versátil y oportuno de la información.

Estas herramientas se refieren por un lado al conocimiento de las técnicas estadísticas por aplicar, y por otro, al manejo rápido y variado de la información.

Es relevante mencionar en este momento que, los datos de las investigaciones, con frecuencia se almacenan en archivos para manejarlos como si provinieran de una investigación de tipo transversal, aún si éstas son en realidad de tipo longitudinal. Los términos transversal y longitudinal se refieren a la evolución del fenómeno en estudio, en el tiempo y/o espacio. Se trata de un estudio transversal cuando, se realiza o realizó sólo una vez la medición o mediciones de las variables del fenómeno; y en el caso contrario, es longitudinal, si se realizan varias mediciones de la(s) variable(s) en las unidades de estudio, en el tiempo⁽¹⁾.

El almacenar de una u otra forma los datos, es usual, dado que con muy poca frecuencia se analizan los datos de manera longitudinal, bien sea por el desconocimiento de los métodos (análisis de mediciones con observaciones repetidas, series de tiempo, análisis de supervivencia, otras) o bien por no contar con software conveniente para preparar los datos para tales análisis. De ahí que comúnmente, los archivos se formen de acuerdo a una investigación transversal, independientemente si ésta lo es.

Si los datos se han archivado para procesarlos como una investigación transversal y en el transcurso del análisis se

(1)Mendez, I. 'El Protocolo de la Investigación.' Ed. Preliminar. Mex. D.F.

repara en que podían ser analizados longitudinalmente, entonces, para procesar los datos con este enfoque, estos deberían estar almacenados de otra manera; lo que pudiera dar opción a capturar los datos para formar los archivos convenientemente (lo cual es muy costoso y tardado); o bien disponer de una herramienta para preparar el archivo de datos como se desee. En este último caso, dicha herramienta es el módulo DM del BMDP.

La problemática planteada con anterioridad es más notoria cuando en un estudio longitudinal se requiere tanto el análisis longitudinal como del transversal, es decir, necesariamente se requiere formar dos diferentes tipos de archivos.

Para entender la importancia del manejo de archivos se presentan algunos ejemplos donde esta necesidad es patente.

1.2 Ejemplos

Los ejemplos que a continuación se mencionan, se refieren a investigaciones que se han analizado en el Laboratorio de Asesoría de Estadística del Proyecto Académico: Especialización en Estadística Aplicada.

1.2.1 Investigación 1. Relación entre Migración y Salud Mental.

En este trabajo se pretendía conocer si la Migración de la gente del campo a la ciudad tiene algún efecto en su salud mental; con una encuesta comparativa observacional prospectiva de tipo transversal, dada la naturaleza del estudio. Observacional, por no interferir el investigador en las mediciones, interfiere sólo en el aspecto de ser captada (no en el pasado) de acuerdo a sus objetivos: prospectiva; y transversal, dado que sólo en una ocasión fueron medidas las variables de interés, es decir no hubo mediciones posteriores de otra época. Es una encuesta comparativa, pues, se "tienen dos o más poblaciones y se pretende comparar algunas variables de ellas, en una ocasión única. Esto se hace con el objeto de contrastar una o varias hipótesis".⁽²⁾

(2) Héndez, I. 'El Protocolo de la Investigación'. Ed. Prelim. Mex. D.F.

Para ello se estudió la asociación entre los factores psico-sociales (migración) y las enfermedades mentales (depresión, ansiedad, y somatizaciones), comparando a una muestra de individuos no migrantes (299) con una muestra de individuos migrantes (290) y un total de preguntas o variables de 595 por sujeto, aproximadamente, para los migrantes y de 481 para los no migrantes.

Cada archivo tiene además variables adicionales que son generadas a partir de otras; por ejemplo, para el diagnóstico de la depresión se requiere medir el conjunto de síntomas: ánimo depresivo, si hay sueño excesivo, pérdida de la memoria, pensamiento lento, falta de apetito; entre otras, así como la duración y edad de inicio de los síntomas.

Para el análisis de este trabajo se requirió de un variado manejo de archivos en donde fueron de gran ayuda algunas de las instrucciones del DM para la creación y unión de archivos y para la presentación gráfica de los datos faltantes a través de un mapa.

1.2.2 Investigación 2. Estimación de Total y Varianza del total estimado, para diseños de Muestreo.

En la práctica es frecuente enfrentar el cálculo de mediciones que reflejen un total de una población en estudio, y la variabilidad que exista en la misma, sobre todo si se trata de una gran población. (Se entenderá por 'gran' población, "cuando la cantidad de unidades con que se cuenta sea demasiado grande para estudiarlas a todas con los recursos humanos, físicos y financieros disponibles"⁽³⁾).

A los Diseños de Muestreo que se presentan, se les calcula directamente el estimador del total y su varianza; para el primer diseño, se usa un muestreo estratificado con selección aleatoria simple sin reemplazo, para obtener el total de ventas de una cadena de tiendas de autoservicio; para el segundo diseño, un

(3) IDEM (2). Pag. 38.

muestreo bietápico con selección con probabilidad proporcional al tamaño con reemplazo en la primera etapa y muestreo aleatorio simple sin reemplazo en la segunda etapa, para obtener el total de latas de sopa en reserva. En ambos diseños son calculados los estimadores con instrucciones del DM.

Los estudios para ambos diseños de muestreo, siguen la pauta de un protocolo de encuesta descriptiva prospectiva transversal.⁽⁴⁾

La solución del problema de estimaciones de totales y sus varianzas en Muestreo probabilístico es de interés, dado el poco software existente; sin embargo dicho problema es fácilmente resuelto con la intervención del DATA MANAGER pues, como se verá en su momento, existen párrafos en el DM que facilitan el cálculo de tales estimaciones.

1.2.3 Investigación 3. Efecto de trasplante de tejido cerebral embrionario en la recuperación funcional de ratas con Hemiplejía Experimental (disminución en la función de movimiento de la mitad del cuerpo).

El propósito de esta investigación, fue contrastar hipótesis en base a varias poblaciones de ratas (estudio comparativo) a las que les fue aplicado determinado tratamiento, en un seguimiento de seis meses, que pudiera causar determinado efecto o resultado en los trasplantes de tejido (estudio experimental, longitudinal).

Protocolo 8: Experimento, "un experimento se caracteriza porque se eligen las variantes del factor causal que se quieren investigar y se asignan las unidades experimentales aleatoriamente a dichas variantes".⁽⁵⁾

En esta investigación experimental, se pretendía evaluar el efecto del trasplante de tejido cerebral embrionario en estado de suspensión, en la recuperación del daño cerebral motor, considerando el modelo de hemiplejía provocado, expresado en términos de recuperación inducida, en comparación con la recuperación pasiva en ratas.

(4) IDEM (3). Pag. 15

(5) IDEM (4). Pag. 87.

En el experimento se consideraron en estudio 24 ratas. A cada una de éstas, se asignó aleatoriamente sólo uno de los cinco tratamientos siguientes: control, lesión falsa, lesión verdadera, lesión verdadera más trasplante falso y lesión verdadera más trasplante verdadero; y estuvieron bajo observación durante tres fases de experimentación en el tiempo que duró la investigación. Ocho variables, con un número diferente de mediciones para cada una de las ratas, fueron observadas en cada fase.

Entre las características de la información dada por los investigadores, se partió de la formación de 5 grupos, uno de ellos con diferente número de ratas (cuatro). En otras palabras, el experimento no era balanceado. Esto se refiere a que el número de individuos o casos no es el mismo en los grupos; por lo tanto los tratamientos no se aplican con la misma frecuencia. La comparabilidad de tales grupos, proporcionaría conclusiones al experimento.

Respecto a esta comparabilidad, es conocido que la comparabilidad de los grupos respecto a los factores de confusión ("factores que no son objeto de estudio y que pueden actuar antagónica y sinérgicamente con las variables consideradas como causa"⁽⁶⁾), se logra, realizando la igualación de atributos o bloques y la asignación aleatoria de las unidades a los grupos de estudio. (Los individuos dentro de un bloque son comparables entre sí, pero no con respecto a los de otro bloque, lo que implica que el grupo de individuos que reciben un tratamiento son comparables con el grupo de los que reciben otro tratamiento. "Esta comparabilidad se refiere a que los grupos resultan homogéneos en lo que respecta a los criterios usados para definir los bloques"⁽⁷⁾).

Así, un bloque se formará cuando, en muestras por compararse, las características de los individuos o casos, difieren (factores de confusión); por lo que se trata de homogeneizar la muestras respecto a los factores de confusión, de tal manera que en cada

(6) IDEM (5), Pag. 198.

(7) IDEM (6), Pag. 203.

bloque estén presentes las variables del factor en estudio. Al querer evaluar el efecto de varios tratamientos sobre todos estos casos; será válido, debido a la formación de los bloques.

En este caso, las ratas en estudio no forman bloques (entendiéndose su definición como en el párrafo anterior) o igualación de atributos, ya que no existen características (variables) que distingan a unas de otras; y por tratarse de un estudio experimental, sólo le son asignados los tratamientos en forma aleatoria, lo que determinará la formación de grupos de ratas.

Respecto al número de mediciones para las variables en cada fase; la cantidad de información que se tiene por rata se refiere a una matriz de dimensión: 24×504 .

Esta investigación requería tanto de un análisis transversal como longitudinal, ello en el manejo de archivo de datos se refiere a un problema de transposición de archivos.

El detalle de la solución de los problemas del manejo de archivos de las investigaciones anteriores se describe en el capítulo 3.

1.3 Objetivos del Trabajo.

Con los diversos párrafos del módulo DM es posible manejar la información desde su forma inicial para crear archivos con el total de datos y dejarlos preparados para un análisis de datos posterior. Con el uso de los archivos de las investigaciones aquí presentadas se mostrará la aplicación de los párrafos.

En una forma general, los objetivos están encaminados a darse cuenta de la necesidad de tal aplicación para la solución de diversos problemas en la Estadística. De esta manera los objetivos que se pretenden alcanzar, son:

Ejemplificar:

01. Cómo se pueden mezclar o juntar renglones de dos o más archivos: MERGE, adición de renglones o muestras. Así también; cómo mezclar las columnas contenidas en un archivo con las columnas de otro: JOIN, adición de columnas.

02. Cómo producir nuevas unidades (ejemplo: familia) a partir de unidades en un archivo (individuos de la familia) haciendo uso del párrafo AGGREGATE.

03. Cómo determinar la forma estructural de los datos y cómo localizar a los datos faltantes, de una manera gráfica y rápida realizando un seguimiento a través del tiempo o del espacio, con la intervención del párrafo MAP.

04. Cómo transponer archivos de datos con el uso de los párrafos PACK y UNPACK.

CAPÍTULO 2

IDEAS GENERALES PARA EL MANEJO DE ARCHIVOS CON EL MÓDULO DATA MANAGER (DM) DEL PAQUETE ESTADÍSTICO BMDP.

2.1 Conceptos afines con el módulo DM del BMDP.

Inicialmente se definirán algunos términos que se utilizarán en el transcurso del trabajo.

Se definirá como registro de información muestral o caso (observación o muestra) al conjunto que contiene a las diversas mediciones realizadas en un elemento de una población, llamadas variables. Un archivo es un conjunto de estos registros y puede representarse de una forma matricial, donde tales registros se refieren a los renglones y las columnas a las variables. De este modo se puede decir que las variables son elementos de dichos registros y un registro será sinónimo de renglón en un archivo.

Conviene hacer notar en este momento, para manejo de terminología y conceptual, la diferencia entre dos tipos de registros: el registro del sistema o línea, que se caracteriza por el tamaño de registro(s) de información muestral donde se guarda físicamente la información; y el registro de información muestral o caso.

A este respecto ocurre que, un registro de información muestral o caso, puede estar descrito en uno o varios registros del sistema, o bien lo contrario, varios casos pueden estar descritos en un registro del sistema.

Para efecto de lenguaje, se llamará al registro de información muestral: caso, observación y en ocasiones solamente registro y al registro del sistema: línea o renglón de archivo.

En el módulo DM existen los archivos del sistema, conocidos como los archivos permanentes (system files), y los generados por un permanente, los archivos temporales (work files):

SFILES (system files) son los archivos de entrada de datos almacenados en el directorio del sistema operativo. Este tipo de archivos son permanentes ya que pueden ser leídos en cualquier momento.

FILES (work files) son los archivos de trabajo temporales internos de DM que existen durante una sesión de este módulo y sobre los cuales operan los párrafos de DM. Estos se crean en la memoria de la computadora a partir de los archivos permanentes.

Es importante comentar que los subarchivos son los archivos derivados de los originales. Un archivo temporal puede ser algo más que el permanente cuando en el programa se han agregado o modificado variables. A veces puede ser de interés almacenar este subarchivo como permanente.

En los párrafos de DM se trabaja con Archivos de Entrada y Archivos de Salida:

En el manejo de DM, es fundamental conocer en cada párrafo, cuál(es) archivo(s) es (son) de entrada y cuál el de salida.

Por lo anterior, vale mencionar que cada párrafo de DM requiere al menos un archivo de entrada.

En particular, READ lee archivos permanentes y cada archivo de entrada requiere un párrafo READ. Los demás párrafos, sólo pueden tener de entrada archivos temporales, aunque no necesariamente sólo uno.

En párrafos diferentes de READ, cuando se utilizan de entrada varios archivos; estos se indican explícitamente. En caso de usar la opción implícita (por omisión) se considera como archivo de entrada, sólo uno: el último temporal de salida generado por el párrafo anterior que a su vez genere un archivo de salida.

Cada párrafo genera sólo un archivo de salida. En particular, SAVE genera un permanente; en los demás párrafos que generan archivos de salida, estos son temporales.

READ es el párrafo necesario para crear los archivos de entrada temporales para los demás párrafos. Esto se ilustra en el siguiente esquema:



Las transacciones de salida y entrada entre archivos temporales se dan de un párrafo a otro en DM, de manera sucesiva. Por ejemplo, el archivo temporal de salida del párrafo READ será el archivo temporal de entrada para cualquier párrafo que lo requiera y acepte.

Es común entonces que en un proceso, un archivo temporal de salida de un párrafo sea el archivo temporal de entrada para un párrafo posterior.

Los archivos temporales de salida son generados por los siguientes párrafos:

READ, SORT, CHECK, TRANSFORM, JOIN,
MERGE, AGGREGATE, PACK, UNPACK y EXTRACT.

y aquellos que no los crean, son:

GROUP, PRINT, MAP, STATISTICS, HISTOGRAM, PLOT, CORRLOT y SAVE.

Como ya ha sido mencionado, SAVE es el único párrafo que genera archivos de salida permanentes; esto es:



Dentro del concepto de los archivos permanentes, que maneja el módulo DM, existen los archivos en código BMDP o no entendibles, y los archivos en ASCII ("datos crudos" o entendibles):

Los archivos BMDP son aquellos que almacenan los datos y la información asociada con la estructura de los datos definidos previamente en un Programa de instrucciones en lenguaje BMDP (PILBMDP). Estos archivos tienen las siguientes ventajas:

- I. Menor espacio para el almacenamiento de información.
- II. Graba características de la información: códigos de categorías, nombres de los códigos, nombres de variables, identificación de los datos faltantes.
- III. Al usar el archivo BMDP se considera la información de la estructura de los datos almacenada; sin embargo si se quiere cambiar algo de dicha estructura, basta declarar en los párrafos convenientes, las instrucciones necesarias.

La creación de los archivos BMDP se da en el PILBMDP en el párrafo SAVE de la siguiente manera:

```
SAVE FILE = ARCHIVO. NEW. SFILE = NARCHIVO. CODE = BMDP. /
```

donde CODE es una instrucción de DM que identifica a un archivo BMDP. La primera vez que se define un archivo SFILE, debe darse NEW. Cuando se crea un archivo BMDP debe especificarse un nombre de archivo permanente (SFILES). Opcionalmente se puede utilizar la instrucción KEEP, para señalar qué variables almacenar.

Cuando se edita este tipo de archivo, los códigos que se observan no son directamente entendibles por el lector.

Los archivos en Código ASCII son aquellos donde los datos almacenados son entendibles al ser editados.

La creación de un archivo de este tipo se realiza con

```
SAVE FILE = ARCHIVO. SFILE = NVARCH. NEW. FORMAT = FREE.
```

(pueden utilizarse otros tipos de formato, por ejemplo el fijo, y también es opcional la instrucción KEEP).

La diferencia entre estos dos tipos de archivo radica, como se ha podido observar, en que, en el archivo BMDP, los datos y su estructura se almacenan aunque no son entendibles, mientras que en un archivo ASCII, sólo se almacenan los datos y son entendibles.

Un archivo temporal puede generarse a partir de un archivo en ASCII o de un archivo en lenguaje BMDP y también puede almacenarse como permanente de esas dos maneras.

Otro tipo de archivo que es manejado por el módulo DM, y que con frecuencia se presentan en la práctica, son los llamados Archivos rectangulares y Archivos no rectangulares:

Los archivos rectangulares son aquellos cuyos registros contienen a las mismas variables tanto en su definición como en cantidad, mientras que los no rectangulares se caracterizarán porque no todos los registros contienen a las mismas variables o bien porque el número de variables entre registros es diferente.

Aquellos registros que dan el carácter de no rectangular al archivo, se nombrarán: tipos de registros.

Una situación donde se tienen estos archivos, con dos tipos de registros, puede ser el estudio sociológico de una población donde se han levantado cuestionarios con diferentes preguntas para hombres y mujeres.

Los archivos rectangulares son de entrada y salida, permanentes provenientes de temporales, y generados de archivos BMDP y ASCII.

Los archivos permanentes de entrada no rectangulares para el párrafo READ, pueden ser sólo los archivos de "datos crudos" (archivos en ASCII); y no pueden crearse permanentes no rectangulares a partir de temporales con el párrafo SAVE.

Algunos párrafos de DM aceptan archivos de entrada no rectangulares. Estos párrafos son:

READ, CHECK, GROUP, EXTRACT, TRANSFORM, AGGREGATE, PACK,
PRINT, STATISTICS, HISTOGRAM, MAP y SAVE.

Aquéllos que no los aceptan como entrada, son:

SORT, JOIN, MERGE, PLOT, CORR PLOT y UNPACK.

La instrucción RECTYPE identificará a los tipos de registro que se deseen usar. Una vez identificados, DM manejará a cada tipo como un conjunto de registros pertenecientes a un archivo rectangular.

Los párrafos que sólo aceptan archivos de entrada rectangulares, pueden aplicarse sólo después de usar el párrafo EXTRACT sobre los archivos no rectangulares, puesto que este párrafo extrae, sin eliminar, los diferentes tipos de registros para formar nuevos archivos rectangulares.

Ejemplo.

Uso de SORT cuando en principio el archivo era no rectangular:

READ	arch.perman. entrada no rect. arch.temp. salida no rect.
------	---

EXTRACT	arch.temp. entrada no rect. arch.temp. salida rect.
---------	--

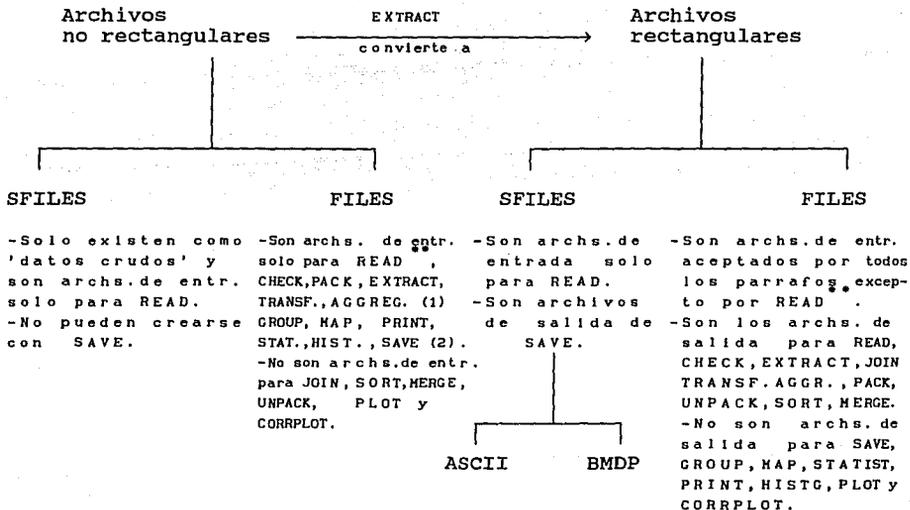
SORT	arch.temp. entrada rect. arch.temp. salida rect.
------	---

(ver párrafo EXTRACT en la sección 2.3)

Generalmente los archivos permanentes son rectangulares, cuando no lo son, a través del temporal pueden ser almacenados como permanentes rectangulares. En el ejemplo anterior, si el archivo de salida del párrafo SORT se deseara almacenar, se lograría con el párrafo SAVE.

Sólo archivos rectangulares y no rectangulares con un tipo de registro (identificado con la instrucción RECTYPE =), pueden ser salvados por SAVE.

Mucho de lo anteriormente mencionado puede esquematizarse como a continuación se muestra:



- (1) crean FILES de salida no rect.
- (2) no crean FILES.

** READ lee solo archivos permanentes: SFILES.
 READ acepta nombrar FILES a un archivo permanente.
 (ver ejemplos en Programa SAVF.DM, p.27 y Programa SERIESTRAS.DM, p. 89)

2.2 Estructura de un programa de instrucciones de lenguaje BMDP (PILBMDP) utilizando el módulo DM.

La estructura de un PILBMDP de DM está constituida por párrafos, los que se delimitan al final por el signo diagonal / ; y en donde estarán contenidas una o varias instrucciones, las cuales indicarán su terminación con el signo punto (.) .

Es conveniente en este momento diferenciar de manera general, la forma como se constituyen las instrucciones en un módulo DM, respecto a cómo se realiza en un PILBMDP en cualquier otro módulo.

El signo / es usado al inicio del párrafo cuando se trata de un módulo del BMDP que no es DM. El párrafo INPUT en módulos diferentes de DM, es equivalente a READ en el módulo DM. Otras diferencias en el uso de los módulos se presentan a continuación:

<u>módulo distinto a DM</u>	<u>módulo DM</u>
/INPUT FILE = .	READ SFILE = . /
/PRINT DATA.	PRINT VAR = ALL. /
/SAVE FILE = .	SAVE FILE =
	SFILE = . /
/END	END /
/FINISH	FINISH /

En el desarrollo del trabajo se comentarán instrucciones de los distintos módulos. Por lo pronto, se presentan en el Cuadro 1, la descripción de algunos párrafos e instrucciones en el módulo DM.

En el CUADRO 1, son Párrafos: READ, SORT, JOIN, UNPACK, SAVE, PRINT, END y FINISH; e Instrucciones: VNAMES, VARIABLES o VAR, KEEP, FORMAT, FILE, SFILE, KEY, NEWFILE, LEVEL, NEWNAMES, NEW y CODE.

CUADRO 1. Algunos párrafos e instrucciones del DM.

READ SFILE =	.	Lee el archivo de entrada permanente (lugar donde los datos están almacenados en disco).
VNAMES =	.	Nombra las variables.
KEEP =	.	Guarda en el archivo temporal, las variables indicadas.
FORMAT =	.	Formato de entrada de las variables.
FILE =	.	Nombra al archivo temporal generado de un archivo en ASCII. ("entendible")
FILE =	.	CODE = Nombra al archivo temporal generado de un archivo BMDP. Este archivo es "no entendible"
<hr/>		
SORT KEY =	./	Variables por ordenar.
<hr/>		
JOIN FILES =	. KEY.	Nombra archivos por juntar en base a variables comunes en los archivos. (Agrega variables).
NEWFILE =	./	Nombra al archivo generado por JOIN.
<hr/>		
UNPACK VARIABLES =	.	Nombra a las variables que se afectarán por el párrafo UNPACK. (cada registro, se descompone en tantos nuevos registros como se indique en Level).
LEVEL =	.	Determina el número de líneas de salida (este número está determinado por las opciones del registro).
NEWNAMES =	./	Da nombre a las nuevas variables.
<hr/>		
SAVE	FILE =	Nombra al archivo temporal de entrada.
	NEW.	Indica que el archivo declarado en FILE, es el primer archivo escrito en un archivo del sistema.
	SFILE =	Da nombre al archivo permanente.
	CODE =	./ Indica que el archivo permanente es almacenado como archivo BMDP.
<hr/>		
PRINT FILE =	.	Nombra específicamente al archivo temporal de entrada,
VAR =	./	Nombra a las variables por imprimir.
<hr/>		
END /		Borra los archivos temporales desde la memoria del programa sin salirse de DM.
<hr/>		
FINISH /		Terminación del programa. Borra todos los archivos temporales desde el sistema (a menos que sean salvados con el párrafo SAVE) antes de que el programa termine y salga del sistema.

2.3 Descripción del Módulo Data Manager.

Data Manager (DM) es un módulo de manipulación de datos, compatible con los otros módulos del paquete BMDP, y permite "poner a punto" archivos que se quieren analizar. Para ello es posible, con DM, ordenar datos, realizar transformaciones de variables, especificar límites de rangos, definir grupos, reestructurar archivos, mezclarlos, producir listados de los datos y gráficas, salvar archivos y convertir archivos no rectangulares en rectangulares.

Los párrafos que constituyen el módulo DM pueden agruparse de acuerdo a sus funciones, como sigue:

Entrada.

Lectura de los archivos de entrada permanentes SFILES con el párrafo READ. Éste genera los archivos temporales FILES, sobre los cuales se realizarán operaciones. Los demás párrafos tienen de entrada al menos un archivo temporal.

Preparación de Archivos.

Se preparan los archivos temporales con el uso de los párrafos: SORT, TRANSFORM, CHECK, y GROUP con lo que se ordenan registros, se crean y transforman variables, se especifican límites de rango y se definen grupos.

Manipulación de Archivos.

Se refiere a la reestructuración del(los) archivo(s), manipulando varios o sobre los registros de uno sólo. Esto se realiza con los párrafos: JOIN, MERGE, AGGREGATE, UNPACK, PACK y EXTRACT.

Reportes de Resultados.

Imprime en un archivo los datos que pudieron ser reestructurados y asimismo, los resultados de reporte de análisis; con la intervención de los párrafos: PRINT, MAP, STATISTICS, HISTOGRAM, PLOT Y CORRPLOT.

Salida.

Pueden salvarse, de forma permanente, los archivos con SAVE, para ser usados en otros programas. Algunos de los demás párrafos tienen como salida archivos temporales. Pueden controlarse: instrucciones de archivos, memoria, mensajes de errores con CONTROL y borrar conjuntos de datos desde la memoria del programa con END y DELETE. Con FINISH borra todos los temporales y sale de DM.

La descripción de los párrafos que a continuación se hace, permitirá cubrir someramente su uso, pero se complementará con las Investigaciones del Capítulo 3.

Antes de esta descripción, es conveniente mencionar la forma que se usa para mandar a proceso un paquete de instrucciones del Data Manager de BMDP.

Con el uso del editor EVA en la VAX 11/750, se puede crear el PILBMDP (programa de instrucciones en lenguaje BMDP), en este caso; el nombre de este programa es INFAGRUP.DM.

El símbolo ' \$ ' denota la permanencia en la Vax. Con el comando ' eva ' ; se edita un archivo, por ejemplo:

```
$eva infagrup.dm      <return>
```

```
----- la pantalla despliega el conjunto de -----
----- instrucciones referidas en el archivo de -----
----- nombre INFAGRUP.DM (ver programa INFAGRUP.DM -----
----- mas adelante). -----
```

El archivo de nombre 'EJERCIC.DAT', es el archivo de datos que usa el conjunto de instrucciones INFAGRUP.DM; también creado en el editor de la Vax.

El conjunto de datos (archivo EJERCIC.DAT), se expresa enseguida:

30	1	1	19	70	144	Bernardo
20	1	1	20	66	132	Javier
5	1	1	21	66	128	Miguel
38	1	1	21	90	138	Benito
29	1	1	22	68	126	Ernesto
15	1	1	23	92	134	Juan
10	1	1	28	90	144	Alfredo
1	1	1	31	62	126	Alberto
36	1	1	34	62	132	Guillerm
21	1	1	38	70	122	Arturo
22	1	2	20	80	136	Samuel
7	1	2	21	68	120	Hugo
39	1	2	21	66	142	Gregorio
12	1	2	22	74	134	Marcos
28	1	2	22	70	120	Francisc
17	1	2	24	76	158	Roberto
32	1	2	26	72	126	Bruno
3	1	2	28	64	128	Felipe
40	1	2	30	70	132	Ricardo
27	1	2	32	68	116	Cristian
37	1	2	35	74	116	Carlos
16	1	2	41	68	112	David
2	2	1	20	78	154	Maria
19	2	1	21	88	156	Juana
9	2	1	22	88	160	Rosa
6	2	1	27	96	265	Catia
23	2	1	33	76	148	America
13	2	1	45	66	148	Susana
14	2	2	19	68	142	Dolores
11	2	2	21	82	140	Pati
18	2	2	21	86	146	Alicia
31	2	2	21	86	144	Nancy
26	2	2	22	80	158	Berta
34	2	2	24	72	142	Barbara
24	2	2	25	78	148	Ana
35	2	2	28	80	138	Pilar
4	2	2	29	96	155	Linda
33	2	2	32	84	136	Margarit
25	2	2	37	76	136	Silvia
8	2	2	42	72	138	Angela

En cada registro de este archivo hay ocho campos para ocho variables que están expresadas de la siguiente manera:

El primer campo corresponde al número de identificación del sujeto (en el programa, a esta variable, se le llama ID); el segundo campo a la variable de nombre SEXO, codificada como 1 para el sexo masculino y 2 para el femenino; en el tercer campo se coloca la variable que denota el "statu quo" fumador del individuo, llamada 'FUMA?', representada entre comillas por la presencia del caracter ' ? ', y con la siguiente codificación: 1 para fumador y 2 para no fumador.

El cuarto campo representa a la edad, con este mismo nombre, para la variable; los dos siguientes: PULSOAEJ, PULSODEJ, son las variables para asignar los valores del pulso antes y después del ejercicio, respectivamente. Finalmente, el nombre del individuo, se representa en dos campos de cuatro caracteres cada uno, pues es la capacidad máxima para variables alfanuméricas, al usar formato libre.

El programa INFAGRUP.DM es enviado a proceso con el siguiente comando:

```
$bmdp dm infagrup.dm s.a <return>
```

donde se invoca al Data Manager del paquete BMDP, con: bmdp dm, enseguida se especifica el nombre del archivo que se manda a proceso (en este caso, infagrup.dm) y finalmente se le da nombre al archivo de salida: s.a, como ejemplo.

Descripción de los Párrafos.

La descripción se apoyará en el programa INFAGRUP.DM, cuyo propósito es obtener por un lado, el total hombres y mujeres que hacen ejercicio, y por otro, sus respectivas estadísticas (media, S.D., valores MIN y MAX) en los grupos de edades: '30omenos', '31a40', '41a50' y 'masde50' en las mediciones del pulso antes y después del ejercicio.

Respecto al contenido del programa INFAGRUP.DM:

Este programa ejemplifica el uso de los párrafos READ, CHECK, SORT, PRINT, GROUP, AGGREGATE, STATISTICS, END y FINISH; así como el uso de las instrucciones FROMF, FROMV del párrafo GROUP cuando se ha ejecutado el párrafo AGGREGATE, por no crear GROUP archivos de salida, concepto que se ahondará más en su momento (ver referencia de los párrafos). Además, el uso del párrafo PRINT, despliega la información que tienen los archivos temporales en diversos momentos de la manipulación de los datos.

Programa INFAGRUP.DM:

```

READ      SFILE = 'EJERCIC.DAT'. (No.1)
          VNAMES = ID, SEXO, 'FUMA?', EDAD, PULSOAEJ, PULSODEJ,
          NOMBRE1, NOMBRE2.
          FORMAT = FREE.
          RLABELS = 7, 8. /
CHECK     MIN = 1, 1, 1, 15, 40, 60. (No.2)
          MAX = 50, 2, 2, 60, 200, 200. /
SORT     KEY = SEXO, 'FUMA?', EDAD, NOMBRE1, NOMBRE2. / (No.3)
PRINT VAR = ALL. FIELDS = 2, 2*1, 2*2, 3, 2*-4. / (No.4)
GROUP    CUTPOINTS(EDAD) = 30, 40, 50. (No.5)
          NAMES(EDAD) = '30omenos', '31a40', '41a50', 'masde50'./
PRINT VAR = ALL. / (No.6)
AGGREGATE WITHIN = SEXO. (No.7)
          'TI HEJER' = N(SEXO).
          MAXCOPY = 25. APPEND.
          NEWFILE = TEMEJER. /
PRINT VAR = ALL. / (No.8)
SORT     FILE = TEMEJER. (No.9)
          KEY = SEXO, EDAD./
PRINT VAR = ALL. / (No.10)
GROUP    FILE = TEMEJER. (No.11)
          FROMF = EJERCIC.
          FROMV = EDAD. /
PRINT VAR = ALL. / (No.12)
STATISTICS WITHIN = SEXO, EDAD. (No.13)
          VARIABLE = PULSOAEJ, PULSODEJ. /
END/ (No.14)
FINISH/

```

Explicación del programa INFAGRUP.DM.

(No.1). El párrafo **READ** permite leer el archivo del sistema de nombre 'ejercic.dat' cuyas variables se nombran con la instrucción VNAMES. Se define la lectura con FORMAT=FREE. En este formato, cada variable de un caso se separará de uno o más blancos.

De las ocho variables, las tres primeras son categóricas, las tres siguientes numéricas, y la séptima y octava "alfanuméricas" (nombrel y nombre2). Estas dos últimas se identifican con la instrucción RLABELS, donde se le indica, que la séptima y octava variable son alfanuméricas.

Del archivo de sistema de nombre ejercic.dat, el párrafo READ crea un archivo temporal de salida, que por omisión, lo nombra como ejercic. El término "omisión" significa que el módulo DM, usa el nombre del archivo de entrada como de salida, al no encontrar una instrucción (FILE= .) que especifique qué nombre asignar.

(No.2). El archivo temporal `ejercic` es el de entrada para el siguiente párrafo: `CHECK`, puesto que nuevamente no se designa alguno de entrada en particular para este párrafo, con la instrucción `FILE=`. Aquí se identifican los datos cuyos valores exceden los límites mínimos o máximos permitidos: para la primer variable (ID) el valor mínimo que puede tener es 1 y el máximo 50. Las variables `sexo` y `'fuma?'`, como valor mínimo 1, como máximo 2, etc. Cuando una variable, no cumple con los límites permitidos, el registro a donde pertenece tal variable, no es tomado en cuenta en párrafos como `STATISTICS`, `HISTOGRAM` (ver su descripción respectiva).

En este caso sólo se revisan las 6 variables no alfanuméricas. El orden en que se expresan los límites, corresponde al de aparición de las variables.

El párrafo `CHECK` crea un archivo temporal de salida, `ejercic`; quien ya contiene la revisión de los límites en las variables y es el archivo temporal de entrada para el siguiente párrafo: `SORT`.

(No.3). En este caso, el párrafo `SORT`, ordena el archivo de nombre `ejercic` en la variable `sexo` (primero a los individuos con código 1 y luego a los de código 2), dentro de `sexo` ordena a los individuos con *"statu quo"* fumador de código 1 y luego a los de código 2; dentro de cada *"statu quo"*, ordena ahora por edad; y dentro de edad, los registros son ordenados por `nombre1` y `nombre2`. El nuevo archivo temporal de salida del párrafo `SORT`, se sobrescribe en el archivo de entrada de este párrafo; entonces el nuevo archivo temporal es `ejercic`, el cual contiene la información de revisión de límites de mínimos y máximos (archivo *'chechado'*) y ordenamiento (archivo ordenado). (El párrafo `SORT` trabaja con el siguiente orden: 0,1,2,...,9 A,B,...,Y,Z).

(No.4). Con tales características, este último archivo constituye el archivo temporal de entrada para el siguiente párrafo: `PRINT`, donde se especifican los campos de impresión para cada variable: Un campo de dos caracteres para la primer variable `id` (denotado por un sólo número: 2, que representa a su vez, dos caracteres);

dos campos, (cada uno de un caracter, denotado por *1) para las variables sexo, 'fuma?'; dos campos, cada uno de dos caracteres (*2) para las variables edad, pulsoaej; un campo de tres caracteres (3) para pulsodej y dos campos alfanuméricos de cuatro caracteres (estos caracteres alfanuméricos se denotan por : -) para las variables nombre1 y nombre2.

Este párrafo no crea archivo temporal de salida, por lo que el archivo de entrada para el siguiente párrafo, que es GROUP, es el archivo temporal de salida del párrafo SORT (No.3).

(No.5). El párrafo GROUP define subpoblaciones de registros llamados grupos. En este caso, se definen intervalos de clase para la variable numérica edad. Con la instrucción CUTPOINTS se determinan los valores o puntos de corte que se relacionarán con los cuatro intervalos o categorías declaradas en NAMES.

El párrafo GROUP no altera archivos temporales por lo que al archivo temporal ejercic, sólo se le adiciona la información de agrupamiento como puede observarse en el párrafo PRINT, No.6: la variable edad está categorizada antes de entrar al párrafo AGGREGATE (No.7).

(No.7). El archivo de entrada para el párrafo AGGREGATE es entonces el archivo temporal ejercic 'chechado', ordenado y con la variable edad agrupada. En este párrafo, para cada categoría de la variable declarada en la instrucción WITHIN (sexo): se contabiliza en la variable creada 'TI_HEJER' el total de hombres y mujeres, con la instrucción N (); la instrucción MAXCOPY permite para cada categoría, una contabilización de 25 registros, (mayor a la definida por omisión en DM que es de 20). El valor de la nueva variable 'TI_HEJER', se graba a la derecha de cada registro original al declarar la instrucción APPEND. Aquí, al archivo temporal de salida, se le nombra Temejer, con la instrucción NEWFILE. Este archivo contiene la información del archivo temporal ejercic (archivo 'chechado', ordenado por sexo, 'fuma?' y edad, sólo que sin agrupación, debido a que el párrafo AGGREGATE no guarda la información agrupada), más la nueva variable 'TI HEJER'.

Esto último se reafirma con la impresión del archivo temejer en el párrafo PRINT (No.8).

Es importante mencionar que si en este momento se le da un PRINT FILE = EJERCIC. VAR = ALL. / , se imprime el archivo ejercic con el agrupamiento, con lo que se confirma que el archivo temporal de nombre ejercic sigue existiendo aunque se definan nuevos archivos, lo que permitirá pueda ser usado por algún otro párrafo que lo requiera (en este caso, en la instrucción FROMF= . del párrafo GROUP, No.11). Con PRINT (No.8), se imprime el archivo temporal de salida (Temejer) del párrafo AGGREGATE.

(No.9) SORT. Sólo se ordenan las variables sexo y edad en Temejer, para obtener posteriormente estadísticas por sexo y edad.

El archivo temporal de salida de este párrafo SORT, contiene ahora este ordenamiento (se aprecia en el párrafo PRINT, No. 10).

En el proceso No.8, el archivo temporal ejercic, pierde la información de agrupamiento en la variable edad, por lo que en este GROUP (No.11) se solicita que el archivo temporal Temejer incluya el agrupamiento mencionado, lo cual se logra con FROMV=EDAD que a su vez se recupera del archivo temporal ejercic, con la instrucción FROMF=EJERCIC. Esto puede apreciarse nuevamente con el párrafo PRINT, No.12.

(No.13). El archivo temporal Temejer que contiene ordenación de la variable sexo y edad agrupada y la nueva variable 'TI_HEJER', es el archivo de entrada para el párrafo STATISTICS.

Aquí, se definen subgrupos con las variables categóricas sexo y edad al declararlas en la instrucción WITHIN, (hombres de '30omenos', hombres de '31a40', etc.) para los cuales se calcularán las estadísticas en las variables numéricas pulsoaej y pulsodej (instrucción VARIABLE).

Es notorio que la variable edad aparece agrupada en los cuadros estadísticos debido a la intervención del párrafo GROUP (No.11) en el archivo Temejer.

Por otro lado, es notable también, que el registro que no cumple con los límites permitidos (ID=6), no es contemplado en el cálculo de las estadísticas para la variable PULSOAEJ.

Enseguida se presentan los cuadros resultantes para las dos variables de pulso.

VARIABLE PULSOAEJ

SEXO	WITHIN EDAD	N	MEAN	S.D.	MIN.	MAX.
1	300menos	16	73.875	9.2799	64.000	92.000
	31a40	5	67.200	5.2154	62.000	74.000
	41a50	1	68.000		68.000	68.000
	TOTAL	22	72.091	8.6954	62.000	92.000
2	300menos	13	82.923	8.2710	68.000	96.000
	31a40	3	78.667	4.6188	76.000	84.000
	41a50	2	69.000	4.2426	66.000	72.000
	TOTAL	18	80.667	8.5130	66.000	96.000
TOTAL		40	75.950	9.5379	62.000	96.000

VARIABLE PULSODEJ

SEXO	WITHIN EDAD	N	MEAN	S.D.	MIN.	MAX.
1	300menos	16	133.88	9.865	120.00	158.00
	31a40	5	122.40	6.841	116.00	132.00
	41a50	1	112.00		112.00	112.00
	TOTAL	22	130.27	10.907	112.00	158.00
2	300menos	12	148.58	7.657	138.00	160.00
	31a40	3	140.00	6.928	136.00	148.00
	41a50	2	143.00	7.071	138.00	148.00
	TOTAL	17	146.41	7.882	136.00	160.00
TOTAL		39	137.31	12.555	112.00	160.00

(No.14). Los párrafos **END** y **FINISH** borran todos los archivos temporales, y **FINISH** permite salir de DM.

Previo a la continuación de la explicación de los párrafos restantes, se detalla un pequeño programa donde se expresa la lectura con el párrafo **READ** de un archivo permanente **BMDP**, al cual se nombra por la expresión **FILE**. Con la instrucción **DIRECTORY**, el programa de nombre **SAVF.DM**, obtiene una descripción de los contenidos del archivo **BMDP** de nombre 'novopuls.sav': nombre de **CODE**, de **LABEL**, nombres de las variables y número de casos.

Programa SAVF.DM:

```
READ FILE = 'NOVOPULS.SAV'.  
  DIRECTORY. /  
END/  
FINISH/
```

Salida del Programa SAVF.DM:

```
=====
READ FILE = 'NOVOPULS.SAV'.
  DIRECTORY. /
=====
----- BMDP FILE DIRECTORY -----
FILE NAME IS
DATA SET NUMBER 1
CODE. . . IS NOVOPULS
CONTENT . IS DATA
LABEL . . IS
NUMBER OF VARIABLES 9
CASE LABELING VARIABLES ARE NOMBRE1 NOMBRE2
VARIABLES ARE
  1 ID 2 SEXO 3 FUMA? 4 EDAD 5 PULSOAEJ
  6 PULSODEJ 7 NOMBRE1 8 NOMBRE2 9 DIFERENC
NUMBER OF DATA RECORDS IS 40
-----
*** NOTE *** NO DATA ARE READ WHEN YOU SPECIFY DIRECTORY.
=====
```

El programa **SAVSF.DM** utiliza, en el párrafo **READ**, la instrucción **SFILE**, con lo que se puede observar el uso indistinto de **FILE** o **SFILE** para denominar al archivo permanente 'novopuls.sav'.

Lo muestra el programa **SAVSF.DM**:

Programa SAVSF.DM:

```
READ SFILE = 'NOVOPULS.SAV'.
  DIRECTORY. /
END/
FINISH/
```

Salida del Programa SAVSF.DM:

```
=====
READ SFILE = 'NOVOPULS.SAV'.
  DIRECTORY. /
=====
```

----- BMDP FILE DIRECTORY -----

```
FILE NAME IS
DATA SET NUMBER 1
CODE. . . IS NOVOPULS
CONTENT . IS DATA
LABEL . . IS
NUMBER OF VARIABLES 9
CASE LABELING VARIABLES ARE NOMBRE1 NOMBRE2
VARIABLES ARE
      1 ID          2 SEXO          3 FUMA?          4 EDAD          5 PULSOAEJ
      6 PULSODEJ    7 NOMBRE1       8 NOMBRE2         9 DIFERENC
NUMBER OF DATA RECORDS IS 40
```

*** NOTE *** NO DATA ARE READ WHEN YOU SPECIFY DIRECTORY.

Este resultado permitiría concluir que el párrafo READ acepte se le nombre a un archivo permanente con la instrucción FILE.

Con el fin de complementar la información de AGGREGATE, y apreciar su potencial; ahora se explica el párrafo TRANSFORM y se hace un comentario sobre sus similitudes y diferencias con AGGREGATE.

TRANSFORM reexpresa los datos, deriva nuevas variables, selecciona casos y ejecuta cálculos complejos, usando sus funciones y operadores. Estas funciones, existen en una gran variedad (véase manual DM del BMDP); y son similares para el párrafo AGGREGATE.

TRANSFORM actúa sobre los datos, dado que usa operadores aritméticos comunes: + , - , * , / , ** , entre otros. Además,

obtiene información sobre varias variables de un registro a la vez; por ejemplo, si interesara sacar el máximo valor sobre las variables $y_1, y_2, y_3, \dots, y_6$:

```
elmayor = MAX (y1, y2, y3, y4, y5, y6).
```

O bien para obtener la media de tres puntuaciones de un registro (funciones de argumento múltiple):

```
promedio = MEAN(puntuac1,puntuac2,puntuac3).
```

Reexpresa datos a través de funciones:

Ejemplo: $\log pe = \text{LOG}(\text{peso}). \quad a = \text{EXP}(-.23 * \text{tiempo}).$

Con la función USE es posible que seleccione registros cuando utiliza variables en combinación con operadores lógicos:

```
USE = edad GT 20 AND sex EQ 1.
```

Asignación condicional de valores:

```
IF (Y EQ 0) THEN (log y = LOG(.01))
```

Creación de nuevas variables, por ejemplo a partir de 4 funciones de distribución (Normal, t, chi-cuadrada y F) y sus inversas.

Ejemplo: Se desea obtener la función de distribución t para $t = 2.05$ con 26 grados de libertad:

```
NEWVAR = TCDF(2.05,26)
```

El programa reporta un valor entre 0 y 1 como el área a la izquierda del valor de t. Aquí, $\text{NEWVAR} = .98$.

Con TRANSFORM, es posible utilizar la instrucción FOR%, que permite generar una lista de nombres sin necesidad de listar cada uno individualmente.

Ejemplo, Suma: SUM(FOR K = 1 to 20. % PREG|K, %)

En este caso, la función SUM totaliza las respuestas para 20 preguntas: PREG1, PREG2, PREG3,...,PREG20. El signo " | " enlaza a la variable PREG con un valor de K. Otra aplicación, se presenta en el programa SERIESFIN.DM , página 93.

Como se ha comentado, TRANSFORM permite obtener valores nuevos de los ya existentes, de igual manera actúa AGGREGATE; sólo que en este párrafo se realiza la ejecución sobre una misma variable (o una columna en varios registros a la vez) como se podrá apreciar en ejemplos posteriores; y TRANSFORM se ejecuta sobre variables de un sólo registro (renglón, caso, observación, muestra).

Para la explicación de algunos párrafos más, estos usan otros programas de referencia.

JOIN mezcla variables distintas de registros similares de diferentes archivos, en base a una o más variables en común en los archivos. Puede decirse que este párrafo mezcla horizontalmente los archivos.

Ejemplo:

```

JOIN      FILES = histor, laborat, pregunt.
          KEY   = id.
          NEWFILE = datcom.
    
```

El párrafo JOIN mezcla los archivos temporales histor, laborat y pregunt en base a la variable en común id. El archivo temporal de salida donde se encuentra la información de los tres archivos se nombra datcom con la instrucción NEWFILE.

La naturaleza de los archivos temporales de entrada y salida, es:

Entrada:

histor	laborat			pregunt		
id ed sex hrs trb	id niv col. wbc	id no.1 no.2 ...no.1				
95 25 1 30	95 205 8.52	95 0 2 5				
96 39 2 40	96 256 15.12	96 5 1 . . . 2				
97 28 1 25	97 195 9.36	97 4 3 5				
.				

95 25 1 30	95 205 8.52	95 0 2 5
96 39 2 40	96 256 15.12	96 5 1 . . . 2
97 28 1 25	97 195 9.36	97 4 3 5
.

95 25 1 30	95 205 8.52	95 0 2 5
96 39 2 40	96 256 15.12	96 5 1 . . . 2
97 28 1 25	97 195 9.36	97 4 3 5
.

Salida: datcom

id	ed	sex	hrs trb	niv col.	wbc	no.1	no.2	...	no.10
95	25	1	30	205	8.52	0	2		5
96	39	2	40	256	15.12	5	1	...	2
97	28	1	25	195	9.36	4	3		5

El uso de JOIN es común cuando los datos son recolectados de diferentes fuentes y en diferentes tiempos.

MERGE mezcla e intercala registros de diferentes archivos, con la opción del uso de una o varias variables en común en los archivos.

En otros términos, se dice que la instrucción **MERGE** combina el final (último registro) de un archivo con el inicio de otro (primer registro). Se pensaría en una mezcla de archivos de tipo vertical cuando no se usa la variable común en los archivos.

Ejemplo. Mezcla de dos archivos temporales de nombre ciudad1, ciudad2, sin el uso de la variable en común, y creación del nuevo archivo temporal dosciud.

```
MERGE      FILES = ciudad1, ciudad2.  
           NEWFILE = dosciud. /
```

Archivo temporal
ciudad1:

id	sexo	fuma?	edad	pulsoaej	pulsodej
30	1	1	19	70	144
20	1	1	20	66	132
5	1	1	21	66	128

Archivo temporal
ciudad2:

3	1	2	28	64	128
39	1	2	21	66	142

Archivo temporal
de salida dosciud:

30	1	1	19	70	144
20	1	1	20	66	132
5	1	1	21	66	128
3	1	2	28	64	128
39	1	2	21	66	142

En dosciud se aprecia la mezcla de los registros de los archivos.

El orden en que se unen los archivos, es el mismo en que fueron declarados en el párrafo MERGE.

En el siguiente programa, de nombre, MERG.DM, puede apreciarse cómo el párrafo en cuestión intercala los registros de los archivos, con el uso de una variable en común. DM supone que esta variable en los archivos de entrada está ordenada ascendentemente, a menos que exista otra orden.

Programa MERG.DM:

```

READ  SFILES = 'ciudad1.dat'.
      FORMAT = FREE.
      VNAMES = ID, SEXO, 'FUMA?', EDAD, PULSOAEJ, PULSODEJ.
      FILE = CD1. /
READ  SFILES = 'ciudad2.dat'.
      FORMAT = FREE.
      VNAMES = ID, SEXO, 'FUMA?', EDAD, PULSOAEJ, PULSODEJ.
      FILE = CD2. /
MERGE  FILES = CD1, CD2.
      KEY = ID.
      NEWFILE = DOSCIU. /
PRINT  FILE = DOSCIU.
      VAR = ALL. / sin la instruc. VAR=ALL, no imprime
END/
FINISH/

```

Arch. permanente ciudad1.dat:

Arch. permanente ciudad2.dat:

5	1	1	21	66	128
20	1	1	20	66	132
30	1	1	19	70	144

3	1	2	28	64	128
39	1	2	21	66	142

Salida del programa MERG.DM:

```
=====
MERGE      FILES = CD1, CD2.
           KEY = ID.
           NEWFILE = DOSCIU. /
=====
```

3 RECORDS READ FROM FILE CD1 2 RECORDS READ FROM FILE CD2

5 RECORDS MERGED ONTO FILE DOSCIU

PATTERN	COUNT	ACTION TAKEN
***	0	FIRST

ID	SEXO	FUMA?	EDAD	PULSOAEJ	PULSODEJ
3.0	1.0	2.0	28.0	64.0	128.0
5.0	1.0	1.0	21.0	66.0	128.0
20.0	1.0	1.0	20.0	66.0	132.0
30.0	1.0	1.0	19.0	70.0	144.0
39.0	1.0	2.0	21.0	66.0	142.0

PACK crea un renglón de archivo o línea a partir de valores representados en columna, o bien crea un renglón con un conjunto de renglones.

Ejemplos:

a) Generación de un renglón a partir de varios, cuando existe un solo caso o individuo, (no es necesaria la instrucción WITHIN):

id	edad	sexo	seq	y
1	35	1	1	44
1	35	1	2	48
1	35	1	3	43
1	35	1	4	46
1	35	1	5	49

↓

```
PACK          VAR = y.
              LEVEL = 5.
              KEEP = id, edad, sexo. /
              NEWNAMES = ymed1 TO ymed5. /
```

↓

id	edad	sexo	ymed1	ymed2	ymed3	ymed4	ymed5
1	35	1	44	48	43	46	49

En este ejemplo, y , es la variable por ser afectada por el párrafo PACK. En LEVEL se define el número de renglones de entrada por arreglar en uno, para el individuo u observación, y en KEEP, se declaran las variables que acompañan a las variables afectadas por PACK.

b) Generación de un renglón a partir de varios, cuando existe más de un individuo o caso, (uso de WITHIN):

Sea un archivo con tres individuos y cada uno con cinco renglones de archivo:

	id	edad	sexo	seq	y
5 renglones	1	35	1	1	44
	1	35	1	2	48
5 renglones	1	35	1	5	49
	2	28	1	1	45
5 renglones	2	28	1	2	23
	2	28	1	5	52
5 renglones	3	30	2	1	39
	3	30	2	2	40
5 renglones	3	30	2	5	47

```

PACK WITHIN = id.
      VAR = y.
      LEVEL = 5.
      KEEP = id, edad, sexo.
      NEWNAMES = ymed1 TO ymed5. /
  
```

id	edad	sexo	ymed1	ymed2	...	ymed5
1	35	1	44	48	...	49
2	28	1	45	23	...	52
3	30	2	39	40	...	47

Según se aprecia se requirió acomodar la variable y, por individuo, esto es, de varios renglones a uno.

Cada individuo con sus mediciones, es un conjunto, el cual se define con la instrucción WITHIN, sobre la variable ID.

Aún se conserva LEVEL=5 : cada individuo consta de cinco renglones para las mediciones de y.

UNPACK crea muchos renglones a partir de uno, es decir, las mediciones de un caso o registro representadas en renglón, las coloca en una columna.

Del mismo ejemplo (inciso a) para PACK; se desean colocar las variables ymed1 a ymed5 del archivo, en posición de columna:

id	edad	sexo	ymed1	ymed2	ymed3	ymed4	ymed5
1	35	1	44	48	43	46	49



```
UNPACK  VARIABLES = ymed1 TO ymed5.
        KEEP      = id, edad, sexo.
        NEWNAME   = id, edad, sexo, seq, y. /
```



id	edad	sexo	seq	y
1	35	1	1	44
1	35	1	1	48
1	35	1	1	43
1	35	1	1	46
1	35	1	5	49

Las variables definidas en la instrucción VARIABLES, del párrafo UNPACK, serán aquellas por arreglar en columna, siguiendo el mismo orden en que estén definidas en el archivo de entrada.

Para cada caso o registro de entrada, el número de variables definidas en VARIABLES será el número de renglones creados en el archivo de salida.

NEWNAMEs lista los nombres de las variables de los renglones de salida y SEQ es una variable que enumera estos renglones, (en este caso, de 1 a 5).

Los párrafos PACK y UNPACK realizan la transposición de archivos; más no en la forma usual conocida en álgebra de matrices, sino como a continuación se muestra.

Ejemplo forma usual:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}, \quad A' = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix}$$

En cambio, DM transpone de la forma siguiente:

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad B'_{(DM)} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{bmatrix}$$

EXTRACT selecciona registros de archivos no rectangulares, selecciona subconjuntos de variables de entrada (reduciendo el ancho del archivo, con la instrucción DELETE) y copia archivos.

Ejemplos.

a) Para copiar un archivo:

```
EXTRACT      FILE = arch1.      NEWFILE = arch2. /
```

donde el archivo temporal de nombre arch1 se copia a uno de nombre arch2. Las variables podrán ser renombradas con NEWNAME.

b) Para hacer un archivo rectangular de uno no rectangular:

Se desea leer un archivo no rectangular y seleccionar dos tipos de registros, para posteriormente juntar información respecto a dos de sus variables, de tal forma de tener un archivo rectangular.

Se presenta el archivo no rectangular de nombre danrect que contiene diferente información en sus registros, es decir, sus variables no son las mismas. Sólo tienen en común a la variable ID_F (identificador de familia) y a la variable ID_C (identificador de ciudad), H e I en la columna TR son los identificadores de los tipos de registros.

Archivo DANRECT:

ID_F	ID_C	N	TR	MF	S	MOD
1.0	1.0	3.0	H			
1.0	1.0		I	1.0	2.0	1978.0
1.0	1.0		I	2.0	1.0	1980.0
1.0	1.0		I	3.0	1.0	1992.0
1.0	2.0	3.0	H			
1.0	2.0		I	1.0	2.0	1978.0
1.0	2.0		I	2.0	1.0	1980.0
1.0	2.0		I	3.0	1.0	1992.0
1.0	3.0	3.0	H			
1.0	3.0		I	1.0	2.0	1978.0
1.0	3.0		I	2.0	1.0	1980.0
1.0	3.0		I	3.0	1.0	1992.0

Los registros con identificador H contienen ID_F, ID_C y N (número de individuos en la familia), mientras que los registros de identificador I, contienen ID_F, ID_C, MF (número de miembro en la familia), S (sexo: masculino 1, femenino 2) y MOD (modelo de auto).

Se pretende formar un nuevo archivo que sea rectangular y donde exista la información de los registros H e I en cada registro; es decir, un archivo como el siguiente:

ID_F	ID_C	MF	S	MOD
1.0	1.0	1.0	2.0	1978
1.0	1.0	2.0	1.0	1980
1.0	1.0	3.0	1.0	1992
1.0	2.0	1.0	2.0	1978
1.0	2.0	2.0	1.0	1980
1.0	2.0	3.0	1.0	1992
1.0	3.0	1.0	2.0	1978
1.0	3.0	2.0	1.0	1980
1.0	3.0	3.0	1.0	1992

Este archivo rectangular, se crea con el programa ANRECT.DM:

Programa ANRECT.DM:

```

READ SFILE = DANRECT.
  RECID = 8.
  RECTYPES = H, I.
  VNAME(H) = ID_F, ID_C, N.
  VNAME(I) = ID_F, ID_C, MF, S, MOD.
  FORMAT = FREE. /
EXTRACT RECTYPE = H.
  NEWFILE = CASA. /
SORT KEY = ID_F, ID_C. /
PRINT VAR = ALL. /
EXTRACT FILE = DANRECT.
  RECTYPE = I.
  NEWFILE = INDI. /
SORT KEY = ID_F, ID_C. /
PRINT VAR = ALL. /
JOIN FILES = CASA, INDI.
  KEY = ID_F, ID_C.
  NEWFILE = TODOS.
  KEEP = ID_F, ID_C, MF, S, MOD.
PRINT VAR = ALL. /
END /
FINISH /

```

Explicación del Programa ANRECT.DM:

El programa ANRECT.DM lee el archivo permanente no rectangular de nombre danrect.

Con la instrucción RECID se le indica la columna donde encontrará los identificadores H e I, los que se declaran con la instrucción RECTYPES= . Los nombres de las variables, por tipo de registro, se declaran con VNAME ()= ; y pueden ser leídas con formato libre.

Aquí, el registro con identificador H tiene tres variables, mientras que el de I tiene cinco.

El primer párrafo EXTRACT selecciona, del archivo temporal de salida del párrafo READ, a aquellos registros con identificador H. Se genera un nuevo archivo temporal de nombre casa, y es el archivo de salida para el párrafo EXTRACT.

El archivo de entrada del párrafo SORT, es rectangular, proveniente del temporal de nombre casa. Entonces, se ordena este archivo rectangular respecto a las variables en común en los dos archivos: ID_F , ID_C; y por omisión, el archivo de salida del párrafo SORT, es: casa.

Con el párrafo PRINT, es posible conocer, hasta este momento la naturaleza del archivo de nombre casa.

Se requiere ahora formar un archivo rectangular para los individuos; para lo cual se procede análogamente en los registros con identificador I, con el segundo párrafo EXTRACT, sobre el archivo temporal danrect.

El archivo temporal de salida de este segundo párrafo EXTRACT, es un archivo rectangular proveniente del temporal de nombre danrect. Al archivo temporal de salida se le denomina indi; el cual también se ordena. Nuevamente, con el párrafo PRINT, se conoce la naturaleza de indi.

Con el párrafo JOIN se juntan los archivos: casa e indi, en

base a las variables en común ID_F e ID_C, y se crea el archivo temporal de nombre todos. Este archivo tiene ya la información respecto a casa e individuos. Se imprimen las variables de KEEP.

Salida del programa ANRECT.DM:

```
=====
READ SFILE = DANRECT.
  RECID = 8.
  RECTYPES = H, I.
  VNAME(H) = ID_F, ID_C, N.
  VNAME(I) = ID_F, ID_C, MF, S, MOD.
  FORMAT = FREE.
  /
=====
```

```
RECORDS H OF FILE DANRECT ARE READ WITH FREE FORMAT.
RECORDS I OF FILE DANRECT ARE READ WITH FREE FORMAT.
12 CASES HAVE BEEN READ. FILE NAME DANRECT NUMBER OF CASES 12
```

RECORD_TYPE	LABELS	V	A	R	I	A	B	L	E	N	A	M	E	S
H		ID_F		ID_C		N								
I		ID_F		ID_C		MF		S						MOD

```
=====
EXTRACT RECTYPE = H.
  NEWFILE = CASA.
  /
=====
```

RECORD_TYPE	LABELS	V	A	R	I	A	B	L	E	N	A	M	E	S
H		ID_F		ID_C		N								
I		ID_F		ID_C		MF		S						MOD

```
OUTPUT FILE:CASA NO. OF CASES: 3
LABEL VARIABLES: VARIABLES: 1 ID_F 2 ID_C 3 N
```

```
=====
SORT KEY = ID_F, ID_C.
INPUT FILE:CASA OUTPUT FILE:CASA NUMBER OF CASES SORTED 3
PRINT VAR = ALL.
  /
=====
```

_ID_F	ID_C	N__
1.0	1.0	3.0
1.0	2.0	3.0
1.0	3.0	3.0

```
=====
EXTRACT FILE = DANRECT.
  RECTYPE = I.
  NEWFILE = INDI.
  /
=====
```

RECORD_TYPE	LABELS	V	A	R	I	A	B	L	E	N	A	M	E	S
H		ID_F		ID_C		N								
I		ID_F		ID_C		MF		S						MOD

```
OUTPUT FILE:INDI NO.OF CASES: 9 LABEL VARIABLES:
VARIABLES: 1 ID_F 2 ID_C 3 MF 4 S 5 MOD
```

```
=====
SORT KEY = ID_F, ID_C.
INPUT FILE:INDI OUTPUT FILE:INDI NUMBER OF CASES SORTED 9
```

PRINT VAR = ALL. /

ID_F	ID_C	MF	S	MOD
1.0	1.0	1.0	2.0	1978.0
1.0	1.0	2.0	1.0	1980.0
1.0	1.0	3.0	1.0	1992.0
1.0	2.0	1.0	2.0	1978.0
1.0	2.0	2.0	1.0	1980.0
1.0	2.0	3.0	1.0	1992.0
1.0	3.0	1.0	2.0	1978.0
1.0	3.0	2.0	1.0	1980.0
1.0	3.0	3.0	1.0	1992.0

JOIN FILES = CASA, INDI.

KEY = ID_F, ID_C.

NEWFILE = TODOS.

KEEP = ID_F, ID_C, MF, S, MOD. /

3 RECORDS READ FROM FILE CASA

9 RECORDS READ FROM FILE INDI

9 RECORDS WRITTEN ONTO FILE TODOS

9 OF THESE RECORDS HAVE USE=1.

FILE NAME:TODOS NO.OF CASES:9 LABEL VARIABLES :

VARIABLES: 1 ID_F 2 ID_C 3 MF 4 S 5 MOD

PRINT VAR =ALL. /

ID_F	ID_C	MF	S	MOD
1.0	1.0	1.0	2.0	1978.0
1.0	1.0	2.0	1.0	1980.0
1.0	1.0	3.0	1.0	1992.0
1.0	2.0	1.0	2.0	1978.0
1.0	2.0	2.0	1.0	1980.0
1.0	2.0	3.0	1.0	1992.0
1.0	3.0	1.0	2.0	1978.0
1.0	3.0	2.0	1.0	1980.0
1.0	3.0	3.0	1.0	1992.0

Existe la opción de no contemplar KEEP en el programa ANRECT.DM, :

JOIN FILES = CASA, INDI.

KEY = ID_F, ID_C.

NEWFILE = TODOS.

KEEP = ID_F, ID_C, MF, S, MOD. /

ID_F	ID_C	N	MF	S	MOD
1.0	1.0	3.0	1.0	2.0	1978.0
1.0	1.0	MISSING	2.0	1.0	1980.0
1.0	1.0	MISSING	3.0	1.0	1992.0
1.0	2.0	3.0	1.0	2.0	1978.0
1.0	2.0	MISSING	2.0	1.0	1980.0
1.0	2.0	MISSING	3.0	1.0	1992.0
1.0	3.0	3.0	1.0	2.0	1978.0
1.0	3.0	MISSING	2.0	1.0	1980.0
1.0	3.0	MISSING	3.0	1.0	1992.0

MAP Aquí se examina la estructura de los registros del archivo, destacando el número y distribución de las observaciones a través de tiempo o del espacio. Este párrafo provee entonces, un mapa de los registros donde cada caracter del mapa representa la presencia de una medición en su eje horizontal.

Es de utilidad en estudios donde hay gran cantidad de datos.

Ejemplos:

Ejemplo 1. Los datos siguientes pertenecen a un estudio experimental, longitudinal, de tres tratamientos clínicos:

Los pacientes fueron asignados aleatoriamente a uno de tres tratamientos: droga 1 (A), droga 2 (B) y placebo (C); y sus efectos se registraron semanalmente durante un año. Si el paciente tuvo una reacción adversa al tratamiento aplicado, se sujetó a un periodo de lavado (D) antes de serle aplicado otro tratamiento.

La información consta de 36 individuos, cada uno de ellos identificado con un número ID no consecutivo, y con distintas semanas de medición; donde se registra el efecto de la droga. Si se piensa que la información está dada por semana y por individuo, puede pensarse en un archivo en donde cada individuo tiene tantos renglones de archivo como semanas de medición registre tal individuo, por lo que la siguiente distribución puede dar idea de cómo está construida la información original:

<u>id</u>	<u>semana</u>	<u>droga</u>
1	0	C
1	1	C
1	2	C
1	3	C
1	5	C
1	.	.
2	52	B
2	1	B
2	.	.
2	35	D
2	.	.
72	0	A
72	.	.
72	47	A

Para una rápida visualización, se supone están ordenados los registros por número de id, no obstante en el siguiente programa de instrucciones donde se ejemplifica el párrafo MAP, se solicita la ordenación de tal variable, ya que en ocasiones los registros no presentan ordenamiento.

La información ha sido resguardada en un archivo BMDP, llamado datdroga, que contiene casi 800 registros.

Las instrucciones con que se obtiene el mapa del archivo datdroga, son:

READ	SFILE = DATDROGA.		
	CODE = ESTDROGA.		/
SORT	KEY = id.	/	/
MAP	WITHIN = id.		
	VARIABLE = droga.		
	TIME = semana.	/	

Explicación del paquete de instrucciones para obtener el mapa del archivo datdrog:

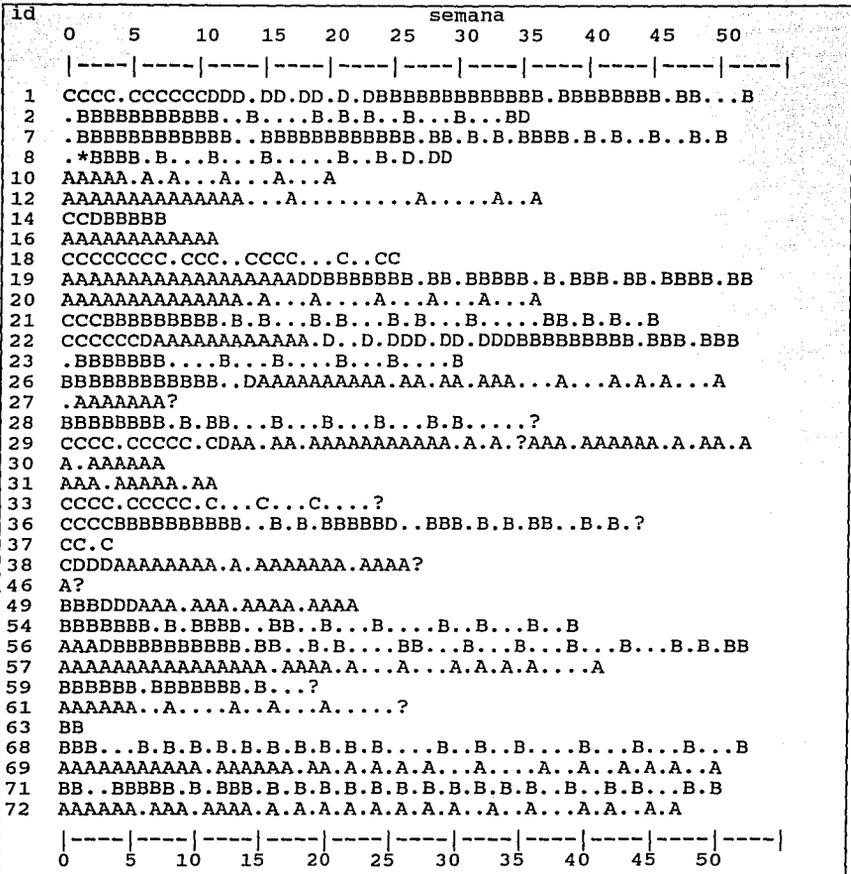
READ lee el archivo BMDP llamado datdroga. SORT lo ordena en base a la variable id.

En el párrafo MAP, cada nivel de cada variable listada en WITHIN, define un renglón del mapa (en este caso, id no tiene niveles, entonces un número id es un renglón del mapa).

La instrucción VARIABLE identifica a la variable de nombre droga, cuyos niveles o categorías se representan con letras en el mapa. En este caso, tales niveles son cuatro: A, B, C y D.

TIME define a la variable semana como el eje horizontal del mapa, lo que permite expresar la numeración de las semanas de cero a cincuenta y cinco.

Mapa para el archivo datdroga:



MINIMUM TIME 0
 MAXIMUM TIME 52

droga	CODE	SYMBOL
	1	A
	2	B
	3	C
	4	D
MISSING VALUE		?
MULTIPLE INCIDENCE		*

NUMERO DE INDIVIDUOS

<u>semana</u>	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>	<u>?</u>	<u>TOTAL</u>	<u>CUMM</u>
0	13	8	10	0	0	31	36<-
1	12	13	9	1	1	36	36
2	13	10	7	2	0	32	34
3	11	10	7	3	0	31	34
4	13	13	2	1	0	29	33
5	12	13	5	1	0	31	33

siguen datos de las semanas de la 6 a la 48

49	1	3	0	0	0	4	10
50	3	3	0	0	0	6	10
51	0	3	0	0	0	3	6
52	1	4	0	0	0	5	5
53	0	0	0	0	0	0	0

-> 36 SUBJECTS HAVE RECORDING AT OR AFTER semana - 0.
 31 SUBJECTS HAVE RECORDING AT semana - 0.
 OF WHICH 13 CODE A, 8 CODE B
 10 CODE C
 0 CODE D
 0 UNKNOWN.

CUMM -NUMBER OF SUBJECTS (WITHIN SETS) WITH RECORDS UP TO THIS TIME OR FURTHER.

TOTAL -REPORTS RECORD COUNTS FOR EACH SPECIFIC TIME POINT.

Explicación del mapa:

El símbolo " ." indica la ausencia del valor de la variable declarada en TIME (ausencia del paciente a la clínica en determinada semana); el signo " ? " denota valores faltantes físicamente o no registrados por alguna otra causa y el asterisco " * " representa a los valores múltiples.

El tiempo mínimo (MINIMUM TIME) se refiere al valor de la variable declarada en TIME donde se registró la primer medición y el máximo al mayor que reportó información.

En una tabla se relacionan los valores de la variable declarada en TIME con los niveles de la variable declarada en la instrucción VARIABLE.

La columna TOTAL sumaliza los reportes de los niveles de las variables declaradas en VARIABLE más los reportes faltantes ("?) y CUMM obtiene el total de caracteres registrados por columna (por semana), ya sea " . " , " * " , " ? " o los símbolos A, B, C, D.

La flecha " <- " indica el valor de la variable declarada en TIME donde hubo el primer registro o medición.

Cada renglón del mapa representa las visitas de cada paciente.

Por ejemplo, al paciente con número id de valor 19 se le aplicó la droga 1 desde la semana cero a la 17, tuvo dos semanas de periodo de lavado (D) y fue aleatorizado a la droga 2, tratamiento que casi concluyó.

Son pocos los pacientes que tienen el mismo comportamiento de visitas a la clínica. Además, es posible deducir de la gráfica que mientras el tiempo de estudio crece, la frecuencia de las visitas clínicas decrece y muchos abandonan los tratamientos.

Otra información que proporciona la gráfica es, conocer el comportamiento de una semana específica. Así, en la semana 3; 11 pacientes fueron tratados con la droga 1, 10 con la droga 2, 7 con el placebo y 3 tuvieron periodo de lavado.

Ejemplo 2. Se presenta ahora, el mapa para el archivo 'ejercic.dat' , usado en el programa INFAGRUP.DM, y donde se solicitan dos variables en la instrucción WITHIN: 'fuma?', sexo:

```
=====
SORT      KEY = 'FUMA?', SEXO. /
=====
```

```
=====
MAP WITHIN = 'FUMA?', SEXO.
      VARIABLE = SEXO.
      TIME = EDAD. /
=====
```

Salida:

MAP FOR FILE EJERCIC

```

FUMA?                                EDAD
                                     20      30      40
                                     15      25      35      45
+-----+-----+-----+-----+
1  ....AA*AA...A..A..A...A
1  .....BBB...B....B.....B
2  .....A**A.A.A.A.A.A....A
2  ....B.*B.BB..BB..B...B...B
+-----+-----+-----+-----+
                                     15      25      35      45
                                     20      30      40

```

MINIMUM TIME 19
MAXIMUM TIME 45

<u>SEXO</u>	<u>CODE</u>	<u>SYMBOL</u>
	1	A
	2	B
MISSING VALUE		?
MULTIPLE INCIDENCE		*

Del mapa se deduce que hay 10 hombres que fuman (A), uno en cada edad 19, 20, 22, 23, 28, 31, 34 y 38 años y una incidencia múltiple en la edad 21.

Existen 6 mujeres fumadoras (B) de edades 20, 21, 22, 27, 33 y 45 años, etc.

Entonces, es posible apreciar mejor la situación de los fumadores y no fumadores hombres y mujeres en relación a su edad.

De la instrucción WITHIN, se puede apreciar que la variable 'fuma?' tiene dos niveles, por lo que define dos renglones del mapa, la variable sexo tiene dos niveles, por lo que define también dos renglones; y la variable sexo declarada en la instrucción VARIABLE, determina la simbología en el mapa: A, B.

STATISTICS reporta estadísticas descriptivas: media, frecuencia, desviación estándar, para variables de entrada, o las generadas en algún párrafo de DM.

Pueden obtenerse estadísticas para subconjuntos de registros como en el siguiente ejemplo.

Ejemplo. El archivo usado es 'ejercic.dat'. El interés es conocer estadísticas de media, tamaño de muestra: n, desviación estándar, mínimo y máximo, para los grupos sexo y "statu quo" fumador en la medición pulsodej; para lo cual se necesita ordenar los datos por los subgrupos formados: mujeres que fuman; mujeres que no fuman, hombres que fuman, hombres que no fuman.

En el párrafo STATISTICS se usa la instrucción WITHIN para definir los subgrupos.

Nota: A la(s) variable(s) declarada(s) en la instrucción WITHIN de los párrafos STATISTICS e HISTOGRAM, o en la instrucción CODES del párrafo GROUP; se le(s) nombra indistintamente: variable o grupo. A los niveles o categorías de una de estas variables, se les denominará: subgrupos.

El conjunto de instrucciones que procesa las estadísticas, es:

```

|SORT                KEY = sexo, fuma?          /
|GROUP CODES(sexo, fuma?) = 1, 2.
|                   NAMES(sexo) = hombres, mujeres.
|                   NAMES(fuma?) = si, no.      /
|STATISTICS         WITHIN = sexo, fuma?.
|                   VARIABLE = pulsodej.       /
  
```



VARIABLE PULSODEJ

WITHIN		N	MEAN	S.D	MIN.	MAX.
SEXO	FUMA?					
HOMBRES	SI	10	132.60	7.545	122.0	144.00
	NO	12	128.33	13.096	112.00	158.00
	TOTAL	22	130.27	10.907	112.00	158.00
MUJERES	SI	6	171.83	45.880	148.00	265.00
	NO	12	143.58	7.128	136.00	158.00
	TOTAL	18	153.00	28.979	136.00	265.00
TOTAL		40	140.50	23.690	112.00	265.00

HISTOGRAM Este párrafo despliega histogramas de variables, por subgrupos definidos por WITHIN. El histograma reporta estadísticas.

Nota: En el cuadro anterior, la cifra de S.D. de mujeres fumadoras, tiene un valor mucho mayor que los obtenidos en la tabla para otros subgrupos. El valor máximo de pulsodej para este subgrupo fue de 265, cuando debe ser de 200 (grabado quizás, en un inicio, incorrectamente).

Antes de solicitar el histograma para este estudio, sería conveniente eliminar este valor con el párrafo CHECK.

Ejemplo. Se usa nuevamente el archivo 'ejercic.dat', el que ha sido ordenado con anterioridad, para poder obtener histogramas de cada subgrupo:

```

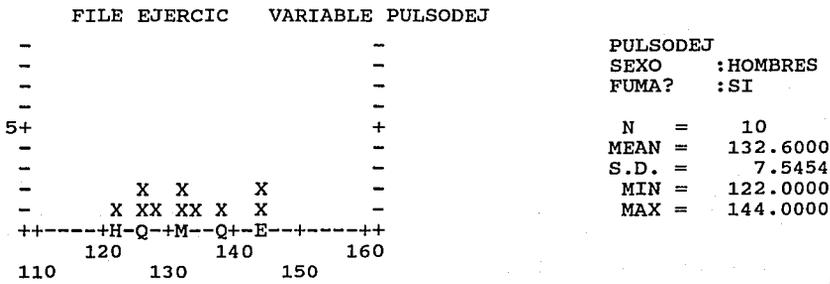
=====
CHECK      MIN = 1, 1, 1, 15, 40, 60.
           MAX = 50, 2, 2, 60, 200, 200. /
=====

=====
GROUP CODES(SEXO) = 1, 2.
        NAMES(SEXO) = HOMBRES, MUJERES.
        CODES('FUMA?') = 1, 2.
        NAMES('FUMA?') = SI, NO. /
=====

=====
HISTOGRAM WITHIN = SEXO, 'FUMA?'.
          VARIABLE = PULSODEJ. /
=====

```

Salida:



El histograma forma una escala automática para el rango de cada variable. Las iniciales M, Q, E, y H en el eje horizontal del histograma, marcan la posición de la media, cuartiles, octiles y sextiles en los datos.

Se obtienen seis histogramas más, uno para cada subgrupo, y para los totales. En el cuadro de abajo, note que el tamaño N en el grupo de mujeres que fuman, antes de seis, es ahora de cinco, debido a que el registro que no cumplió con los límites requeridos, en el párrafo CHECK, no es contemplado para cálculos posteriores.

Se presentan sólo las estadísticas de los histogramas de los subgrupos restantes:

PULSODEJ
SEXO :HOMBRES
FUMA? :NO
N = 12
MEAN = 128.3333
S.D. = 13.0964
MIN = 112.0000
MAX = 158.0000

PULSODEJ
SEXO :HOMBRES
FUMA? : TOTAL
N = 22
MEAN = 130.2727
S.D. = 10.9073
MIN = 112.0000
MAX = 158.0000

PULSODEJ
SEXO :MUJERES
FUMA? :SI
N = 5
MEAN = 153.2000
S.D. = 5.2154
MIN = 148.0000
MAX = 160.0000

PULSODEJ
SEXO :MUJERES
FUMA? :NO
N = 12
MEAN = 143.5833
S.D. = 7.1282
MIN = 136.0000
MAX = 158.0000

PULSODEJ
SEXO :MUJERES
FUMA? : TOTAL
N = 17
MEAN = 146.4118
S.D. = 7.8824
MIN = 136.0000
MAX = 160.0000

PULSODEJ
SEXO : TOTAL
FUMA? : TOTAL
N = 39
MEAN = 137.3077
S.D. = 12.5555
MIN = 112.0000
MAX = 160.0000

SAVE Con la aplicación de este párrafo, de todo archivo temporal se genera un archivo permanente. Pueden salvarse registros resultados del JOIN, MERGE, AGGREGATE, PACK, UNPACK y TRANSFORM; operaciones reestructuradas de un archivo BMDP, un archivo Binario, o un archivo en ASCII.

La instrucción `FORMAT` determina el tipo de archivo (fortran, binario) por salvar. Con esta instrucción, es posible escribir un archivo de datos en ASCII, ya sea para corregir datos, obtener un reporte de los mismos donde se incluyan variables transformadas o nuevas variables; o bien para usar estos valores en otro software.

Ejemplos.

a) Para salvar un archivo en ASCII, se especifica el siguiente tipo de formato:

```
SAVE      FILE = ejercic.  
          NEW.  
          SFILE = 'novopuls.sav'.  
          FORMAT = '(1X, F3.0, 2F2.0, F3.0, 2F4.0, 2A4)'. /
```

b) Para salvar un archivo BMDP se usa la instrucción `CODE`:

```
SAVE      FILE = ejercic.  
          NEW.  
          SFILE = 'novopuls.sav'.  
          CODE = novopuls. /
```

`FILE` nombra al archivo temporal de entrada para el párrafo `SAVE`; si se omite, se salva el archivo temporal generado por el párrafo previo. `SFILE` da nombre al archivo de salida del párrafo `SAVE`, generándolo como archivo permanente BMDP (novopuls.sav) identificado por la instrucción `CODE` con el nombre novopuls.

`NEW` indica que 'novopuls.sav' es el primer archivo escrito en el sistema. Si existen en el sistema otros archivos BMDP que no sea de interés su permanencia, pueden ser borrados al declarar precisamente `NEW`; o bien, pueden ser resguardadas otras versiones del mismo programa, si se omite tal instrucción.

En el siguiente ejemplo se observa cómo la información de otros párrafos se va incorporando en un archivo BMDP:

```

READ   SFILE = 'EJERCIC.DAT'.
       VNAMES = ID, SEXO, 'FUMA?', EDAD, PULSOAEJ, PULSODEJ,
       NOMBRE1, NOMBRE2.
       FORMAT = FREE.
       RLABELS = 7, 8.
SORT   KEY = SEXO, 'FUMA?', EDAD. /
TRANSFORM IF (PULSODEJ EQ 265) THEN PULSODEJ = 165.
       DIFERENC = PULSODEJ-PULSOAEJ. /
GROUP  NAMES (SEXO)= HOMBRES, MUJERES.
       NAMES ('FUMA?')= SI, NO.
       CODES (SEXO, 'FUMA?') = 1, 2.
       NAMES (EDAD) = MENOSD25, '25-34', MASDE34.
CUTPOINTS (EDAD) = 24, 34. /
SAVE   FILE = EJERCIC.
       NEW.
       CODE = NOVOPULS.
       SFILE = 'NOVOPULS.SAV'. /
FINISH /

```



```

=====
SAVE   FILE = EJERCIC.
       NEW.
       CODE = NOVOPULS.
       SFILE = 'NOVOPULS.SAV'. /
=====
DATAMANAGER FILE : EJERCIC
WRITTEN AS FILE : NOVOPULS.SAV           BMDPFILE CODE : NOVOPULS
CONTENT : DATA LABEL : VARIABLES SAVED :
  1 ID          2 SEXO          3 FUMA?    4 EDAD     5 PULSOAEJ
  6 PULSODEJ    7 NOMBRE1       8 NOMBRE2  9 DIFERENC
BMDP FILE HAS BEEN COMPLETED. FILE NAME IS NOVOPULS.SAV
NUMBER OF CASES WRITTEN TO FILE          40

```

El párrafo SAVE salva el archivo temporal de salida del párrafo TRANSFORM, que hasta este momento incluye las ocho variables leídas donde se establece permanentemente el valor de 165 en lugar de 265 para la variable pulsodej, y la nueva variable diferenc. Lo salva entonces como archivo BMDP en un archivo del sistema de nombre 'novopuls.sav'.

CONTROL Otro potencial del Data Manager consiste en seleccionar la forma interactiva o batch de un programa de ejecución, cambiar el tamaño de la memoria y la longitud de una línea, especificar dónde se graban los macro-archivos de instrucciones, obtener diagnósticos cuando un programa con error es suspendido y controlar el número de errores tolerable. Todas estas características se encuentran en el Párrafo CONTROL.

DELETE borra uno o más archivos temporales desde la memoria y libera espacio en la misma.

CAPÍTULO 3.

APLICACIONES EN EL DATA MANAGER.

Con el propósito de percibir el potencial del módulo Data Manager del BMDP, se describirán a continuación tres investigaciones, donde se aplican diversos conceptos de DM.

3.1 Investigación 1.

Relación entre Migración y Salud Mental (Cuevas, 1990)

Esta investigación fue desarrollada bajo el carácter de estudio observacional, prospectivo, transversal y comparativo; para conocer la asociación entre las variables: factores psicosociales y enfermedades mentales, que a juicio del investigador, arrojaría un conocimiento respecto al objetivo por conocer: si la migración de los individuos del campo a la ciudad repercute en la salud mental de los migrantes.

Para esto, extrajo una muestra del campo: 299 individuos no migrantes, a los que se les aplicó un cuestionario de 481 preguntas; para compararla con la que obtuvo de la ciudad: 290 individuos migrantes, a los que se les aplicó un cuestionario de 595 preguntas.

El desarrollo del análisis requirió de la creación de un archivo de datos para analizar conjuntamente a los 290 individuos migrantes y 299 individuos no migrantes. Para ello hubo necesidad de mezclar archivos para individuos migrantes y archivos para no migrantes, estableciendo igual dimensión en los archivos (el mismo número de variables y las mismas), corregir algunos datos por problemas en la captura, crear variables adicionales y conocer su comportamiento estructural.

3.1.1 Creación y mezcla de archivos.

Uno de los procesos de esta investigación requirió crear archivos para los individuos migrantes y archivos para los individuos no migrantes. Se requiere del DM para formarlos:

archivos de datos
para migrantes:

MIGP1
MIGP2
MIGP3

archivos de datos para
no migrantes:

NMIGP1
NMIGP2
NMIGP3

Enseguida se unieron los archivos para migrantes: MIGP1 con MIGP2 y MIGP3, a través del programa JUNTAMIG.DM:

Programa	JUNTAMIG.DM	(módulo DM)
READ	SFILE = MIGP1.	
READ	SFILE = MIGP2.	
READ	SFILE = MIGP3./	
MERGE	FILES = MIGP1, MIGP2 , MIGP3.	
	NEWFILE= MIGT./	
SAVE	SFILE=MIGT. FORMAT = FREE./	
FINISH/		

Lee tres archivos de entrada permanentes: MIGP1, MIGP2 y MIGP3.

El párrafo MERGE mezcla los registros de los archivos temporales (igualmente nombrados, por omisión, que los permanentes) MIGP1, MIGP2 y MIGP3; y forma el nuevo archivo temporal MIGT. Este temporal de salida MIGT es el archivo temporal de entrada para el párrafo SAVE, el cual es salvado con el mismo nombre: MIGT, y en ASCII.

De igual manera se procede para los no migrantes y en este caso el archivo resultante pudiera ser NMIGT.

Como se ha observado, lo realizado hasta este momento se refiere a adicionar información muestral por individuo.

Entonces, para cada clase de individuos se creó un archivo; MIGT para migrantes, NMIGT para no migrantes.

La investigación requirió crear un solo archivo a partir de MIGT y NMIGT. Esto fue posible al dimensionar en igual tamaño, a ambos archivos (a 410 variables); ya que MIGT tenía 506 y NMIGT 437.

Se requirió este último proceso antes de realizar la mezcla de MIGT y NMIGT.

Primeramente, el programa MIG291.1D, dimensionó a 410 variables a los archivos MIGT y NMIGT:

Programa MIG291.1D (módulo 1D de BMDP)

```

/ BMDP 1D
/ INPUT FILE = MIGT.      (archivo de migrantes, 506 variables)
  VARIABLES = 506.
  FORMAT='(i3,i1,7x,2(i2),65x,/,55x,i2,3(i1),3(i2),3(i1),
5(i2),i1,/,80(i1),/,51(i1),i2,i1,i2,i1,i1,i2,20(i1),/,57(i1),
i3,i1,i2,i1,i1,i2,i3(i1),/,45(i1),i3,i2,4(i1),i2,4(i1),2(i2),
16(i1),/,61(i1),i2,i3,i1,i1,i2,8(i1),2x,/,54x,14(i1),4(i3),/,
6(i3),3(i1),11(i2),i1,i3(i2),10x,/,5x,22(i3),i1,4(i2),/,17(i2),
2(i1),7(i3),23x)'.
/VARIABLES NAMES=ident,.... continúan variables .....,pr246.
/TRANSFORM      ocact=ocact1*10+ocact2.
/ SAVE KEEP=ident,... continúan más variables ... ,dsdstythy.
FILE=MIG410.CODE=RED.NEW.
/PRINT DATA.LINESIZE=80.
/END
```

El programa BMDP 1D describe datos.

Se menciona el archivo de entrada de datos permanente en ASCII de nombre MIGT, el número de variables a usar, 506, su formato, el nombre de ellas, y se usa una transformación para corrección de datos y nuevas variables.

MIG410 es salvado como archivo BMDP (con 410 variables, KEEP) y desplegado su contenido con PRINT.

El mismo proceso se realizó para el archivo de no migrantes de 437 variables: NMIGT. Se nombró NMIG410, al archivo de no migrantes, cuyo contenido se sintetizó también a 410 variables.

Ahora, los archivos MIG410 y NMIG410, podrán ser mezclados con el programa JUNTA.DM:

Programa JUNTA.DM

```

READ      SFILE = MIG410 . CODE=RED./
READ      SFILE = NMIG410 . CODE=RED./
MERGE     FILES = MIG410, NMIG410.
          NEWFILE= DOSMT./
SAVE      SFILE=DOSMT.CODE=JUNTOS.NEW./
FINISH/

```

El programa JUNTA.DM, lee los archivos BMDP de nombre MIG410 y NMIG410. De los temporales creados con el párrafo READ, se mezclan los registros de un archivo con los registros del otro; y al temporal creado por MERGE, se le llama DOSMT. Este archivo es el temporal de entrada para el párrafo SAVE, donde se salva como archivo BMDP.

3.1.2 Mapa de datos faltantes.

Resulta claro la necesidad de conocer en un momento dado el comportamiento de los datos, debido a que existen dos grupos grandes de estudio: 290 individuos migrantes, 299 individuos no migrantes; y con un gran número de variables; más de 400.

Con el propósito de ejemplificar, se trabajará con 10 individuos migrantes (10 casos).

Dentro de los 10 casos, dos son individuos hombres y ocho mujeres.

Programa MAP1.DAT

```

READ SFILE = CPAMENVI. FORMAT = FREE.
VNAME = SEX ,MIGR , NTSINTS , NTSINTA , NTSINTD.
FILE = CPAMENVI. CASES = 10. /
SORT KEY = SEX. /
MAP WITHIN = SEX.
VARIABLE = MIGR.
TIME = NTSINTS . /
PRINT VAR = ALL. /
END /
FINISH /

```

El programa MAP1.DAT, lee el archivo de entrada CPAMENVI con formato libre, para diez casos. Define nombres a cinco variables: SEX, MIGR, etc., y da al archivo temporal de salida el mismo nombre: CPAMENVI. Ordena este archivo en base a la variable SEX.

Con el párrafo MAP se obtendrá el mapa cuyos renglones estarán definidos por la variable declarada en la instrucción WITHIN según sus niveles existentes; en este caso, un renglón para hombres, y otro para mujeres. Además, se relaciona el comportamiento del grupo de hombres y mujeres migrantes (VARIABLE=MIGR) en relación al número total de síntomas de somatizaciones (TIME=NTSINTS, que va de 0 a 15, en el eje horizontal). Ver mapa para el archivo de CPAMENVI.

Salida de MAP1.DAT

```

=====
MAPA PARA EL ARCHIVO CPAMENVI
(MAP FOR FILE CPAMENVI)

SEX          5          10          NTSINTS
0          +-----+-----+
1          ..A...A
2          ...AA*A.A..A.A
          +-----+-----+
0          5          10

```

TABLA 1

<u>NTSINTS</u>	<u>NUMBER OF SUBJECTS</u>			
	<u>A</u>	<u>?</u>	<u>TOTAL</u>	<u>CUMM</u>
0	0	0	0	2
1	0	0	0	2
2	1	0	1	2<-
3	1	0	1	2
4	1	0	1	2
5	2	0	2	2
6	1	0	1	2
7	1	0	1	2
8	1	0	1	1
9	0	0	0	1
10	0	0	0	1
11	1	0	1	1
12	0	0	0	1
13	1	0	1	1
14	0	0	0	0

El símbolo " ? ", representa el número de valores faltantes.

MINIMUM TIME 2
 MAXIMUM TIME 13

MIGR	CODE	SYMBOL
	2	A
	MISSING VALUE	?
	MULTIPLE INCIDENCE	*

-> 2 SUBJECTS HAVE RECORDING AT OR AFTER NTSINTS = 2.

1 SUBJECTS HAVE RECORDING AT NTSINTS = 2.
 OF WHICH 1 CODE A 0 UNKNOWN.

CUMM -NUMBER OF SUBJECTS (WITHIN SETS) WITH RECORDS UP TO
 THIS TIME OR FURTHER.

TOTAL -REPORTS RECORD COUNTS FOR EACH SPECIFIC TIME POINT.

=====
 PRINT VAR = ALL. /
 =====

TABLA 2

SEX	MIGR	NTSINTS	NTSINTA	NTSINTD
1.0	2.0	7.0	5.0	1.0
1.0	2.0	2.0	2.0	3.0
2.0	2.0	5.0	0.0	0.0
2.0	2.0	8.0	8.0	1.0
2.0	2.0	11.0	2.0	3.0
2.0	2.0	13.0	11.0	6.0
2.0	2.0	3.0	5.0	5.0
2.0	2.0	6.0	21.0	15.0
2.0	2.0	4.0	0.0	0.0
2.0	2.0	5.0	2.0	1.0

Explicación Salida MAP1.DAT:

Se presentó un valor mínimo para NTSINTS de 2 y máximo de 13.

El grupo de los migrantes (los únicos estudiados en este caso en particular) se identifica con el número 2 en el archivo CPAMENVI. Existen ausencias (.) a cierto número de NTSINTS y no hay valores faltantes (?). En cambio, sí hay una incidencia múltiple en NTSINTS=5, la cual se desglosa en la tabla 1 y 2.

En la TABLA 1 se relacionan los valores de NTSINTS (0 a 14) con los niveles de la variable MIGR (en este caso un nivel, el símbolo A).

Se deduce también que para cero número total de síntomas de somatizaciones (NTSINTS=0) hubo cero incidencias de migrantes (A=0), o bien, hay cero individuos migrantes con cero número total de síntomas de somatizaciones. Existen entonces, cero ocurrencias de migrantes en un número total de síntomas de 0,1,9,10,12 y 14; una ocurrencia de migrantes en un número total de síntomas de 2,3,4,6,7,8,11 y 13; dos ocurrencias de migrantes en un número total de 5 y cero valores faltantes para todos los valores de NTSINTS (? = 0).

Del mapa para el archivo CPAMENVI se puede apreciar que:

-- En un número de 2 y 7 NTSINTS, el grupo de hombres (SEX=1) tuvo incidencia; mientras que en el grupo de mujeres (SEX=2), se registró en un número de 3, 4, 6, 8, 11, 13 NTSINTS y en NTSINTS=5, la incidencia fue múltiple.

-- El grupo de hombres registró ausencias en 0, 1, 3, 4, 5 y 6 NTSINTS.

-- El total de ocho letras A's, más la incidencia múltiple, que equivale a dos letras A's más; hace un total de 10 letras A's, lo que corresponde a los 10 casos tratados en el programa MAP1.DAT.

-- Se citó que de los 10 casos, dos son individuos hombres y ocho mujeres. Uno de los individuos hombres tuvo 2 NTSINTS, y el otro 7, en total 9 NTSINTS; y los individuos restantes, las 8 mujeres, tuvieron un total de 55 NTSINTS: dos de ellas con 5 NTSINTS para cada una y seis de ellas con 3, 4, 6, 8, 11 y 13 NTSINTS.

De aquí puede concluirse que dos hombres tienen 9 síntomas de somatización, y ocho mujeres 55; lo que en promedio resulta 4.5 síntomas de somatización en los hombres, mientras que en las mujeres existe un promedio de 6.8; lo que indica a su vez un mayor grado de somatizaciones en las mujeres (conclusión no generalizada puesto que sólo intervienen diez casos, pero que sin embargo puede coincidir con la realidad).

Con PRINT VAR = ALL imprime todos los valores de las variables declaradas en VNAMES, en el cuadro llamado TABLA 2.

3.2 Investigación 2.

Estimación para Diseños de Muestreo.

Anotaciones acerca de muestreo probabilístico:

Los elementos de una población son las unidades elementales cuya información permitirá inferir acerca de la población.

Una unidad de muestreo puede ser un elemento o bien un conjunto de elementos (conglomerado). En un muestreo de conglomerados una unidad de muestreo será precisamente el conglomerado, a diferencia de un muestreo de elementos.

Cuando se submuestran conglomerados resulta el muestreo polietápico, donde la selección de los elementos resulta de la selección de las unidades de muestreo de dos o más etapas. Cualquier selección mejorará si la población es dividida en subpoblaciones diferentes. Estas subpoblaciones son precisamente los estratos.

En un muestreo polietápico se usa una jerarquía de unidades de muestreo o conglomerados, de tal manera que el elemento pertenezca únicamente a una unidad de muestreo en cada etapa. Cada unidad de primera etapa (unidades primarias) se divide en unidades de segunda etapa (unidades secundarias), hasta llegar a las unidades últimas de muestreo.

El muestreo bietápico es un caso del muestreo de conglomerados en dos etapas y dentro de este tipo de muestreo puede darse el muestreo estratificado.

Para mayor referencia de estos conceptos, ver Leslie Kish, "Muestreo de Encuestas" cap.10.

Cuando se lleva a cabo un esquema de muestreo el investigador se enfrenta con problemas desde cómo seleccionar su información hasta cómo procesarla.

En relación al procesamiento, el DATA MANAGER muestra una solución rápida a dicho problema; lo cual se verá ejemplificando las estimaciones de total y su varianza estimada para dos diseños de muestreo, uno Estratificado y otro Bietápico.

3.2.1 Muestreo estratificado, con selección aleatoria simple sin reemplazo.

(KOJI, A. 1987)

Problema en DM : Agregación de información por estrato.

Se desea conocer el total de ventas de una cadena de tiendas en toda la ciudad.

El total de tiendas está dividido en tres zonas: 23 clase alta, 55 clase media y 49 clase baja.

Por un muestreo estratificado aleatorio se midió el total de ventas en cada una de las tiendas seleccionadas y dentro de cada estrato se hizo un muestreo aleatorio simple sin reemplazo: 5 en la clase alta, 16 en la clase media y 14 en la baja. (Los datos están en términos de millones de pesos).

En seguida, la información se expresa en términos de la notación que se usa para los cálculos de los estimadores.

La información que se tiene es:

$L = 3$: Total de estratos

$N = 127$: Total de unidades en la población: $\sum_{h=1}^L N_h = N$

N_h : Total de unidades en el h-ésimo estrato:

$$N_1 = 23$$

$$N_2 = 55$$

$$N_3 = 49$$

$$N_1 + N_2 + N_3 = 127$$

$n = 35$: Total de unidades en la muestra: $\sum_{h=1}^L n_h = n$

n_h : Tamaño de muestra en el h-ésimo estrato:

$$n_1 = 5 \qquad n_2 = 16 \qquad n_3 = 14$$

$$n_1 + n_2 + n_3 = 35$$

$y_{h,i}$: Venta total en la i-ésima tienda muestra del h-ésimo estrato

(nota: el termino $y_{h,i}$ es el que se expresa en los siguientes estratos, como $y_{h,i}$, $h=1,3$; $i=5, i=16, i=14$).

Estrato 1	Estrato 2	Estrato 3
$y_{1,1} = 85.4$	$y_{2,1} = 74.2$	$y_{3,1} = 72.2$
$y_{1,2} = 94.2$	$y_{2,2} = 72.7$	$y_{3,2} = 64.1$
$y_{1,3} = 103.7$	$y_{2,3} = 89.2$	$y_{3,3} = 59.7$
$y_{1,4} = 74.2$	$y_{2,4} = 109.7$	$y_{3,4} = 47.2$
$y_{1,5} = 93.2$	$y_{2,5} = 97.2$	$y_{3,5} = 84.1$
	$y_{2,6} = 74.2$	$y_{3,6} = 92.1$
	$y_{2,7} = 67.7$	$y_{3,7} = 44.1$
	$y_{2,8} = 84.2$	$y_{3,8} = 51.1$
		$y_{3,9} = 52.7$
		$y_{3,10} = 57.2$
		$y_{3,11} = 57.4$
		$y_{3,12} = 91.2$
		$y_{3,13} = 59.2$
		$y_{3,14} = 63.6$

La información por determinar es:

W_h : Proporción de la Población en el h-ésimo estrato
o factor de ponderación. $= \frac{N_h}{N}$

$$W_1 = \frac{N_1}{N} = \frac{23}{127} = 0.181102, \text{ etc.}$$

\bar{y}_h : Promedio del total de ventas por tienda en el h-ésimo estrato (o media muestral por estrato)

$$= \frac{\sum_{i=1}^{n_h} Y_{h_i}}{n_h} = \frac{Y_h}{n_h}$$

(donde $y_{h.} = \sum_{i=1}^{n_h} Y_{h_i}$ es el total muestral de ventas en el estrato h)

$$\bar{y}_1 = \frac{Y_{1,1} + Y_{1,2} + Y_{1,3} + \dots + Y_{1,5}}{n_1} = \frac{85.4 + \dots + 93.2}{5} = 90.14 \text{ , etc.}$$

$s_{y_h}^2$: Varianza muestral del total de ventas por tienda en el h-ésimo estrato.

$$= \frac{\sum_{i=1}^{n_h} (Y_{h_i} - \bar{y}_h)^2}{n_h - 1}$$

$$s_{y_1}^2 = \frac{(Y_{1,1} - \bar{y}_1)^2 + (Y_{1,2} - \bar{y}_1)^2 + \dots + (Y_{1,5} - \bar{y}_1)^2}{n_1 - 1}$$

$$= \frac{(85.4 - 90.14)^2 + \dots + (93.2 - 90.14)^2}{4} = 121.5671 \text{ , etc.}$$

$$\hat{Y} : \text{Estimador del total poblacional} = \sum_{h=1}^L \frac{\sum_{i=1}^{n_h} Y_{h_i} (N_h)}{n_h} = \sum_{h=1}^L \bar{y}_h N_h =$$

$$= 90.14(23) + 79.21(55) + 63.99(49) = 9,564.3$$

(a N_h/n_h se le llama factor de expansión).

$\hat{V}(\hat{Y})$: Estimador de la varianza del estimador del total

$$= \sum_{h=1}^L N_h^2 \left(1 - \frac{n_h}{N_h}\right) \left(\frac{s_{y_h}^2}{n_h}\right)$$

$$= N_1^2 \left(1 - \frac{n_1}{N_1}\right) \left(\frac{s_{y_1}^2}{n_1}\right) + \dots + N_3^2 \left(1 - \frac{n_3}{N_3}\right) \left(\frac{s_{y_3}^2}{n_3}\right)$$

$$= 23^2 \left(1 - \frac{5}{23}\right) \left(\frac{121.57}{5}\right) + 55^2 \left(1 - \frac{16}{55}\right) \left(\frac{187.4}{16}\right) + 49^2 \left(1 - \frac{14}{49}\right) \left(\frac{238.8}{14}\right) = 64,444$$

$\left(\frac{n_h}{N_h}\right)$ es la fracción de muestreo)

Estos valores pueden resumirse en la tabla siguiente para compararlos con los obtenidos por el programa ESTRAT.DM.

Tabla: Datos para un muestreo estratificado y estimaciones del total y la varianza

h	1	2	3
N_h	23	55	49
W_h	0.181102	0.433071	0.385827
n_{hh}	5	16	14
\bar{Y}_h	90.14	79.21	63.99
$s_{y_h}^2$	121.57	187.41	238.81
\hat{Y}	= 9,564.3		
$\hat{V}(\hat{Y})$	= 64,444.48		

El programa ESTRAT.DM es el conjunto de instrucciones que procesa y obtiene los mismos resultados para \bar{Y}_h , $s_{y_h}^2$, \hat{Y} y $\hat{V}(\hat{Y})$ llamados ' \wedge PYH', ' \wedge V(\wedge YH)', ' \wedge Y' y ' \wedge V(\wedge Y)' respectivamente.

Sea y_{hi} (notación que será equivalente a y_{h1} , mencionada anteriormente) la venta total en la i -ésima tienda muestra del h -ésimo estrato; H designa al estrato, n_{hh} (equivalente a n_h) designa al tamaño de muestra en el h -ésimo estrato y NH (equivalente a N_h) al total de unidades en el h -ésimo estrato; todo para efecto del uso del programa ESTRAT.DM que en seguida se menciona. Además; la notación:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad \text{es equivalente a } \text{'^PYH'} = \text{MEAN}(y_{hi}).$$

$$s_{y_h}^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} \quad \text{equivalente a } \text{'^V(^YH)'} = \text{SD}(y_{hi})^{**2}.$$

$$\hat{Y} = \sum_{h=1}^L \bar{y}_h N_h \quad \text{es equivalente a } \text{'^Y'} = \text{SUM}(\text{'^PYH'} * NH)$$

$$\hat{V}(\hat{Y}) = \sum_{h=1}^L N_h^2 \left(1 - \frac{n_h}{N_h}\right) \left(\frac{s_{y_h}^2}{n_h}\right) \quad \text{es equivalente a}$$

$$\text{'^V^Y'} = \text{SUM}(NH ** 2 * (1 - n_{hh}/NH) * (\text{'^V(^YH)'} / n_{hh})).$$

Programa ESTRAT.DM

```
#BMDP DM /**** ESTIMADORES POR ESTRATO: '^PYH' Y '^V(^YH)'\*****
#BMDP DM /:**** ESTIMADORES PARA EL TOTAL: '^Y' Y '^V^Y'\*****
```

```
READ SFILE = AMEST1.            FORMAT = FREE.
VARIABLES = H , yhi. /
AGGREGATE FILE = AMEST1. WITHIN = H. APPEND.
          nhh = N(H).
          '^PYH' = MEAN(yhi).
          '^V(^YH)' = SD(yhi)**2./
TRANSFORM IF (KASE EQ 5 ) THEN    NH = 23.
          IF (KASE EQ 6 ) THEN    NH = 55.
```

```

IF (KASE EQ 22) THEN NH = 49.
IF (KASE NE 5 AND KASE NE 6 AND KASE NE 22) THEN USE=-1.
AUX=1./
AGGREGATE FILE = AMEST1. WITHIN = AUX . APPEND.
'^Y' = SUM( '^PYH'*NH ).
'^V(Y)' = SUM( NH ** 2 *( 1-nhh/NH ) * ( '^V(^YH)'/ nhh ) )./
SAVE SFILE=RESULTA. DELETE = yHi. FORMAT='(6(x,F7.1),2(x,F10.3))'./
PRINT LINEIZE = 80.
VAR =H,NH,'^PYH','^V(^YH)',nhh,AUX,'^Y','^V(Y)'. /
END /
FINISH /

```

Nota: Lo que se encuentra a la derecha del signo #, se refiere a un comentario dentro del programa.

Archivo de entrada de datos permanente del sistema:

AMEST1.DAT

H	yHi				
1	85.4	2	74.2	3	72.2
1	94.2	2	72.7	3	64.1
1	103.7	2	89.2	3	59.7
1	74.2	2	109.7	3	47.2
1	93.2	2	97.2	3	84.1
		2	74.2	3	92.1
		2	67.7	3	44.1
		2	84.2	3	51.1
		2	75.2	3	52.7
		2	79.4	3	57.2
		2	72.1	3	57.4
		2	92.1	3	91.2
		2	52.1	3	59.2
		2	72.4	3	63.6
		2	67.8		
		2	87.2		

sigue estrato 2

sigue estrato 3

Descripción del programa ESTRAT.DM :

Como puede observarse, el programa tiene como entrada la información de las ventas totales (y_{h_i}) de las tiendas muestra (n_h) en tres estratos (H). Para estimar el total de ventas poblacional en la cadena de tiendas de la ciudad; se obtienen los promedios y las varianzas de las ventas totales por estrato \bar{y}_h y $s_{y_h}^2$ respectivamente, con los cuales se llegará a los estimadores totales (\hat{Y} , $\hat{V}(\hat{Y})$).

Si la información se presenta como lo muestra el archivo de nombre AMEST1, se requiere de un párrafo del módulo DM que sume por columna los valores y_{Hi} (esta notación es la del programa ESTRAT.DM) en cada estrato, para obtener primeramente la media y varianza por estrato. Esto lo resuelve el primer párrafo AGGREGATE.

Enseguida, es conveniente formar un grupo de registros que contenga una sola vez los estimadores por estrato, para sumarizar nuevamente por columna y estimar los totales poblacionales.

Tal grupo de registros se forma con el párrafo TRANSFORM, seleccionando los casos deseados; para después ser aplicado otro párrafo AGGREGATE sobre este último grupo de registros y calcular los estimadores para el total poblacional.

Explicación del programa ESTRAT.DM:

Lee el archivo visible permanente de datos del sistema de nombre AMEST1 con formato libre. Se declaran dos variables a leer (VNAME=). En este caso los registros ya se encuentran ordenados en la variable H; si no fuera así, se requeriría del párrafo SORT.

El párrafo READ formó un archivo temporal del mismo nombre (AMEST1) sobre el cual se ejecutaron los párrafos AGGREGATE, TRANSFORM y AGGREGATE, funcionando como archivo de entrada y salida las veces que fueron necesarias: por ejemplo, el archivo temporal de salida del párrafo READ, fue archivo de entrada para el primer párrafo AGGREGATE; una vez ejecutado AGGREGATE se formó otro temporal del mismo nombre (AMEST1) y fue archivo de salida para este párrafo y de entrada para el párrafo TRANSFORM. Así se continúa el procedimiento de entrada y salida del archivo AMEST1 para los siguientes párrafos.

El primer párrafo AGGREGATE obtuvo la media y la varianza para cada estrato ('^PYH' y '^V(YH)' respectivamente). En el párrafo TRANSFORM se asigna a los casos 5, 6 y 22 su número de tamaño de

muestra respectivo, con las instrucciones condicionales IF...THEN...

Con la instrucción IF ... THEN ... USE = - 1, asigna este valor a los casos que no son 5, 6 y 22, y selecciona entonces precisamente los casos 5, 6 y 22; que tienen la media y varianza por estrato.

Se definió la variable AUX y se le asignó el valor de 1, para formar un grupo con los casos 5, 6 y 22. Esta variable se declaró en WITHIN; para el segundo párrafo AGGREGATE que permite obtener los estimadores totales ('^Y' y '^V(^Y)').

Ahora, el archivo temporal de salida del segundo párrafo AGGREGATE de nombre AMEST1, contiene las variables originales y las variables creadas por los párrafos usados. Este archivo es el de entrada para el párrafo SAVE, donde se define con el nombre de RESULTA a las ocho variables salvadas. Entonces, este archivo es ya permanente y está en ASCII (entendible). Por esta última característica, una de las pruebas de este proceso se hizo con FORMAT=FREE para el párrafo SAVE; lo que hubiera funcionado si todas las variables salvadas cumplieran con un rango no mayor de ochos caracteres.

Precisamente el valor de '^V(^Y)' tiene nueve posiciones (64445.500) por lo que en la salida se reportaban asteriscos en la posición de este valor. Se optó entonces por el formato de salida fijo con el que se imprimen las 8 variables declaradas en la instrucción VAR del párrafo PRINT.

Como observación es conveniente mencionar que el uso de la instrucción FORMAT debe de incluir las comillas que delimitan los paréntesis (donde a su vez se especifica el formato de las variables), es decir, FORMAT = '()'. De otra manera, el programa presenta error.

La siguiente estructura de AMEST1 se encuentra antes de ejecutarse el párrafo TRANSFORM.

Estructura de AMEST1 (antes de ejecutarse el párrafo TRANSFORM y después del primer párrafo AGGREGATE):

H	yHi	NH	^YH	^V(^YH)	
1.0	85.4	23.0	90.1	121.6	
1.0	94.2	23.0	90.1	121.6	
1.0	103.7	23.0	90.1	121.6	
1.0	74.2	23.0	90.1	121.6	
1.0	93.2	23.0	90.1	121.6	caso 5
2.0	74.2	55.0	79.2	187.4	caso 6
2.0	72.7	55.0	79.2	187.4	
2.0	89.2	55.0	79.2	187.4	
2.0	109.7	55.0	79.2	187.4	
2.0	97.2	55.0	79.2	187.4	
2.0	74.2	55.0	79.2	187.4	
2.0	67.7	55.0	79.2	187.4	
2.0	84.2	55.0	79.2	187.4	
2.0	75.2	55.0	79.2	187.4	
2.0	79.4	55.0	79.2	187.4	
2.0	72.1	55.0	79.2	187.4	
2.0	92.1	55.0	79.2	187.4	
2.0	52.1	55.0	79.2	187.4	
2.0	72.4	55.0	79.2	187.4	
2.0	67.8	55.0	79.2	187.4	
2.0	87.2	55.0	79.2	187.4	
3.0	72.2	49.0	64.0	238.8	caso 22
3.0	64.1	49.0	64.0	238.8	
3.0	59.7	49.0	64.0	238.8	
3.0	47.2	49.0	64.0	238.8	
3.0	84.1	49.0	64.0	238.8	
3.0	92.1	49.0	64.0	238.8	
3.0	44.1	49.0	64.0	238.8	
3.0	51.1	49.0	64.0	238.8	
3.0	52.7	49.0	64.0	238.8	
3.0	57.2	49.0	64.0	238.8	
3.0	57.4	49.0	64.0	238.8	
3.0	91.2	49.0	64.0	238.8	
3.0	59.2	49.0	64.0	238.8	
3.0	63.6	49.0	64.0	238.8	

Salida del proceso de nombre ESTRAT.DM:

```
=====
#BMDP DM /**** ESTIMADORES POR ESTRATO : '^PYH' Y '^V(^YH)' ****
#BMDP DM /:**** ESTIMADORES PARA EL TOTAL : '^Y' Y '^V(^Y)' ****
READ SFILE = AMEST1. FORMAT = FREE. VNames= H, yHi. /
=====
```

FILE NAME:AMEST1 NO.OF CASES:35 LABEL VARIABLES:VARIABLES:1 H 2 yHi

```

=====
AGGREGATE FILE = AMEST1. WITHIN = H. APPEND.
      nhh = N(H).
      '^PYH' = MEAN(yHi).
      '^V(^YH)' = SD(yHi)**2./
=====

```

*** WARNING *** THE NAME AMEST1 IS ALREADY USED TO NAME A WORK FILE.
 THE OUTPUT FROM THIS PARAGRAPH WILL REPLACE THE OLD WORK FILE.

```

NUMBER OF RECORDS PROCESSED      :   35
NUMBER OF CASES RETAINED         :   35
NUMBER OF CASES WITH USE GT ZERO :   35
FILE NAME:AMEST1                 NO.OF CASES:35      LABEL VARIABLES :
VARIABLES:  1 H      2 yHi      3 nhh      4 ^PYH      5 ^V(^YH)
=====

```

```

TRANSFORM IF (KASE EQ 5 ) THEN NH = 23.
IF (KASE EQ 6 ) THEN NH = 55.
IF (KASE EQ 22) THEN NH = 49.
IF (KASE NE 5 AND KASE NE 6 AND KASE NE 22) THEN USE=-1.
AUX=1./
=====

```

```

NUMBER OF CASES PROCESSED      :   35
NUMBER OF CASES RETAINED         :    3
NUMBER OF CASES WITH USE GT ZERO :    3
FILE NAME: AMEST1              NO. OF CASES : 3      LABEL VARIABLES:
VARIABLES:  1 H      2 yHi      3 nhh      4 ^PYH      5 ^V(^YH)      6 NH      7 AUX
=====

```

```

AGGREGATE FILE = AMEST1. WITHIN = AUX . APPEND.
      '^Y' = SUM( '^PYH'*NH ).
      '^V(^Y)' = SUM( NH ** 2 *( 1-nhh/NH ) * ( '^V(^YH)'/ nhh ) )./
=====

```

*** WARNING *** THE NAME AMEST1 IS ALREADY USED TO NAME A WORK FILE.
 THE OUTPUT FROM THIS PARAGRAPH WILL REPLACE THE OLD WORK FILE.

```

NUMBER OF RECORDS PROCESSED      :    3
NUMBER OF CASES RETAINED         :    3
NUMBER OF CASES WITH USE GT ZERO :    3
FILE NAME : AMEST1             NO. OF CASES : 3      LABEL VARIABLES :
VARIABLES:1 H  2 yHi  3 nhh  4 ^PYH  5 ^V(^YH)  6 NH  7 AUX  8 ^Y
          9 ^V(^Y)
=====

```

```

SAVE SFIL=RESULTA. DELETE = yHi. FORMAT='(6(x,F7.1),2(x,F10.3))'./
=====

```

```

DATAMANAGER FILE: AMEST1          WRITTEN AS FILE: RESULTA
OUTPUT FORMAT:USER'S
VARIABLES SAVED:1 H  2 nhh  3 ^PYH  4 ^V(^YH)  5 NH  6 AUX  7 ^Y
                8 ^V(^Y)
=====

```

```

USER FILE HAS BEEN COMPLETED.
FILE NAME IS RESULTA
=====

```

```

PRINT LINEZISE=80. VAR =H,NH,'^PYH','^V(^YH)',nhh,AUX,'^Y','^V(^Y)'./
=====

```

H	NH	^PYH	^V(^YH)	nhh	AUX	^Y	^V(^Y)
1.0	23.0	90.1	121.6	5.0	1.0	9565.6	64445.5
2.0	55.0	79.2	187.4	16.0	1.0	9565.6	64445.5
3.0	49.0	64.0	238.8	14.0	1.0	9565.6	64445.5

3.2.2 Muestreo Bietápico con selección con probabilidad proporcional al tamaño con reemplazo en la primera etapa y con muestreo aleatorio simple sin reemplazo en la segunda etapa.

(TREJO, B. 1985)

Problema en DM: Agregación de información por unidades de muestreo.

Se desea estimar el total de latas de sopas que tienen en reserva 150 supermercados que están distribuidos en 15 distritos.

Son seleccionados 6 distritos (selección de la primera etapa: 5,8,10,11,12 y 15).

Dentro del número de supermercados existentes en los 6 distritos se obtiene a la vez una muestra de 4 supermercados (segunda etapa) para cada unidad seleccionada de la primera etapa.

Finalmente se registra el número de latas en reserva (y_{1j}) dentro de los cuatro supermercados muestreados.

El problema radica entonces en estimar valores totales poblacionales para los quince distritos (unidades primarias) a partir de los valores estimados en las últimas unidades seleccionadas (unidades secundarias).

Por lo que se necesita obtener a partir de los promedios muestrales en la segunda etapa (\bar{y}_1), el valor total para cada unidad primaria (\hat{Y}_1); quienes proporcionarán a su vez los estimadores totales poblacionales (\hat{Y} y $\hat{V}(\hat{Y})$).

UNIDADES PRIMAR: distrito		UNIDADES SECUND: supermercado		NUMERO DE LATAS EN RESERVA Y_{ij}			
DISTRITO	total de sup. en el i-distr. M_i	supermerca dos selecc. en la mues.	superm. 1	sup. 2	superm 3	super 4	
			$j=1, m_i$				
1	8	m_i					
2	12						
3	4						
4	20		Y_{11}	Y_{12}	Y_{13}	Y_{14}	
5	14	4	2	3	2	5	
6	6						
7	5		Y_{21}	Y_{22}	Y_{23}	Y_{24}	
8	12	4	4	8	9	7	
9	5		Y_{31}	Y_{32}	Y_{33}	Y_{34}	
10	24	4	1	1	6	4	
11	6	4	Y_{41}	Y_{42}	Y_{43}	Y_{44}	
			4	3	3	6	
12	8	4	Y_{51}	Y_{52}	Y_{53}	Y_{54}	
13	4		1	4	8	7	
14	9		Y_{61}	Y_{62}	Y_{63}	Y_{64}	
15	13	4	2	5	5	4	

$N = 15$ $150 = M$ $24 = m$

Selección primera etapa: 6 distritos (5, 8, 10, 11, 12 y 15).

Selección segunda etapa: 4 supermercados

Se continúa con la referencia de los datos en términos de la notación que se usa:

y_{ij} : valor de la característica de interés (número de latas) de la i-ésima unidad primaria (distrito) y la j-ésima unidad secundaria (supermercado) en la muestra.

$$i = 1, \dots, n; \quad j = 1, \dots, m_i$$

$$n = 6, m_i = 4 \quad (Y_{11}=2, Y_{12}=3, Y_{13}=2, Y_{14}=5, Y_{15}=4, \dots, Y_{64}=4)$$

por ejemplo,

($y_{11}=2$, se refiere a: 2 latas en el primer supermercado seleccionado del primer distrito seleccionado)

($Y_{12}=3$, se refiere a: 3 latas en el segundo supermercado seleccionado del primer distrito seleccionado), etc.

N : Total de unidades primarias en la población = 15

M : Total de unidades secundarias en la población

$$= \sum_{i=1}^N M_i = M_1 + M_2 + \dots + M_{15} = 8 + 12 + \dots + 13 = 150$$

M_i : Total de unidades secundarias de la i -ésima unidad primaria

n : Total de unidades primarias en la muestra = 6

m_i : Total de unidades secundarias muestreadas en cada unidad primaria, en la segunda etapa.

$$i = 1, \dots, n ; m_1 = m_2 = \dots = m_6 = 4.$$

m : Total de unidades secundarias en la muestra

$$= \sum_{i=1}^n m_i = m_1 + m_2 + \dots + m_6 = 4 + 4 + \dots + 4 = 24$$

p_i : Probabilidad de la i -ésima unidad primaria en ser seleccionada

$$= M_i / M.$$

Por determinar:

\bar{Y}_i : Media muestral de la característica de interés en la i -ésima

$$\text{unidad primaria} = \frac{\sum_{j=1}^{m_i} Y_{ij}}{m_i} = \frac{Y_{i.}}{m_i} \quad (i = 1, \dots, n)$$

$$\begin{aligned} \bar{Y}_1 &= \frac{\sum_{j=1}^{m_1} Y_{1j}}{m_1} = \frac{Y_{1.}}{m_1} = \frac{Y_{11} + Y_{12} + Y_{13} + \dots + Y_{1m_1}}{m_1} = \\ &= \frac{Y_{11} + Y_{12} + Y_{13} + Y_{14}}{m_1} = \frac{2+3+2+5}{4} = 3 \text{ etc.} \end{aligned}$$

\hat{Y}_1 : Total de la característica de interés en la i -ésima unidad
 unidad primaria = $M_1(\bar{Y}_1)$, $\bar{Y}_1 = \hat{Y}_1 / M_1$

$$\hat{Y}_1 = M_1(\bar{Y}_1) = 14(3) = 42 , \quad \text{etc.}$$

Hasta aquí, los promedios muestrales \bar{Y}_1 de cada unidad
 primaria; proporcionaron su estimador total: \hat{Y}_1 .

Ahora, el Estimador del total poblacional \hat{Y} ; está determinado
 por:

$$\begin{aligned} \hat{Y} &= \frac{M}{n} \sum_{i=1}^n \frac{\sum_{j=1}^{m_i} Y_{ij}}{m_i} = \frac{M}{n} \sum_{i=1}^n \bar{Y}_i = \frac{M}{n} (\bar{Y}_1 + \bar{Y}_2 + \dots + \bar{Y}_n) \\ &= \frac{150}{6} (3+7+3+4+5+4) = 650 . \end{aligned}$$

En términos de probabilidades, si $M_i = (M)(p_i)$; \hat{Y} , se puede
 expresar como:

$$\hat{Y} = \frac{M}{n} \sum_{i=1}^n \bar{Y}_i = \frac{M}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{M_i} = \frac{M}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{(M)(p_i)} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{p_i} ,$$

Sea $z_i = \frac{\hat{Y}_i}{p_i}$; entonces,

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n z_i = \bar{z} ; \quad \text{que equivale a MEAN}(z_i) \text{ en Data Manager.}$$

$\hat{V}(\hat{Y})$: Varianza estimada del estimador total poblacional =

$$\hat{V}(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{Y}_i}{p_i} - \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{p_i} \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{Y}_i}{p_i} - \hat{Y} \right)^2$$

$$= \frac{1}{6(5)} \left[\left(\frac{42}{14/150} - \frac{1}{6} \left(\frac{42}{14/150} + \frac{52}{13/150} + \dots + \frac{24}{6/150} \right) \right)^2 + \dots + \left(\frac{24(150)}{6} - 650 \right)^2 \right]$$

$$= 8,500 .$$

O bien, en términos de $z_1 = \frac{\hat{Y}_1}{p_1}$. De un resultado anterior: $\hat{Y} = \bar{z}$,

$$\hat{V}(\hat{Y}) = \hat{V}(\bar{z}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{Y}_1}{p_1} - \hat{Y} \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^n (z_1 - \bar{z})^2 = \frac{1}{n} s_{z_1}^2 = s_{\frac{z}{z}}^2$$

El término $\frac{s_z^2}{z}$ es la función SEM(z_1) del módulo DM.

Los valores obtenidos pueden resumirse en la tabla C.1D para su mejor comparación con los obtenidos por el programa CONG.LO:

Tabla C.1D

i	1	2	3	4	5	6
M_1	14	12	24	6	8	13
m_1	4	4	4	4	4	4
\bar{Y}_1	3	7	3	4	5	4
\hat{Y}_1	42	84	72	24	40	52
\hat{Y}	= 650					
$\hat{V}(\hat{Y})$	= 8,500					

El programa CONG.LO es el conjunto de instrucciones que procesa y obtiene los mismos resultados para \hat{Y}_1 , p_1 , z_1 , \hat{Y} y $\hat{V}(\hat{Y})$ representadas en el programa por las variables: '^Yi', pi, zi, '^Y' y '^V(^Y)' respectivamente. m_1 se denota como mii.

Programa CONG.LO:

```
#BMDP DM / : ***** CALCULO DE pi y zi *****
#***** ESTIMADORES PARA EL TOTAL : '^Y' y '^V(^Y)' ****
READ SFILE = MC. VARIABLES = 6. FORMAT = FREE.
VNames = NDIS, M1, y11, y12, y13, y14. /
TRANSFORM M = 150. mii = 4.
```

```

'^Yi' = Mi * (SUM(yi1, yi2, yi3, yi4) / mii).
pi = Mi / M.      zi = '^Yi' / pi.      AUX=1. /
AGGREGATE FILE = MC. WITHIN = AUX. APPEND.
'^Y' = MEAN(zi).
'^V(^Y)' = SEM(zi) ** 2. /
PRINT LINESIZE=80. VAR = ALL. /
END /
FINISH /

```

Archivo de datos, permanente y de entrada; de nombre MC:

5	14	2	3	2	5
8	12	4	8	9	7
10	24	1	1	6	4
11	6	4	3	3	6
12	8	1	4	8	7
15	13	2	5	5	4

Justificación del Programa CONG.LO:

Semejante al problema 3.2.1, con la información disponible, se desean obtener los estimadores poblacionales \hat{Y} y $\hat{V}(\hat{Y})$, a partir de subunidades muestreadas. Los estimadores \hat{Y} y $\hat{V}(\hat{Y})$ se obtienen calculando primeramente \hat{Y}_i , p_i y z_i , como ya se ha mencionado.

La información se presenta como en los cuadros explicativos, es decir; cada distrito seleccionado (selección 1a. etapa), el número de supermercados en cada uno de estos distritos y el número de latas en reserva de las unidades secundarias muestreadas y_{ij} , se presentan como renglones en el archivo de datos MC.

Entonces se requiere del párrafo TRANSFORM para que sume los valores y_{ij} renglón por renglón y obtener \bar{y}_i , que al multiplicar por M_i resulte \hat{Y}_i . Además, en este párrafo se definen los valores constantes M y mii . En este caso, existe el mismo tamaño para todas las unidades secundarias muestreadas ($mii=4$). El solicitar en este párrafo los valores para las

expresiones p_i y z_i , facilita el cálculo para el siguiente párrafo.

El definir un valor común en los registros (AUX = 1) permitirá formar un grupo que será identificado en el párrafo AGGREGATE para sumar por columna los promedios muestrales de las unidades primarias.

Explicación del Programa CONG.LO:

Lee el archivo permanente de datos de nombre MC con formato libre que contiene 6 variables, cuyos nombres se dan en la instrucción VNAMEs.

En el párrafo TRANSFORM se definen los valores para M y mii.

La instrucción SUM calcula los promedios muestrales, y el operador (*) los totales en la i-ésima unidad primaria: '^Yi'. Se crean las variables p_i , z_i y AUX=1. Esta última variable será la variable WITHIN en el párrafo AGGREGATE que forma el grupo de registros sobre el cual se ejecutarán por columna las instrucciones MEAN y SEM.

El párrafo PRINT permite visualizar los valores de todas las variables.

Archivo Temporal de nombre MC, de salida y de entrada, después del párrafo TRANSFORM:

NDIS	Mi	yi1	yi2	yi3	yi4	M	mii	^Yi
5.00	14.00	2.00	3.00	2.00	5.00	150.00	4.00	42.00
8.00	12.00	4.00	8.00	9.00	7.00	150.00	4.00	84.00
10.00	24.00	1.00	1.00	6.00	4.00	150.00	4.00	72.00
11.00	6.00	4.00	3.00	3.00	6.00	150.00	4.00	24.00
12.00	8.00	1.00	4.00	8.00	7.00	150.00	4.00	40.00
15.00	13.00	2.00	5.00	5.00	4.00	150.00	4.00	52.00

pi	zi	AUX
0.09	450.00	1.00
0.08	1050.00	1.00
0.16	450.00	1.00
0.04	600.00	1.00
0.05	750.00	1.00
0.09	600.00	1.00

Sobre este archivo se ejecuta el párrafo AGGREGATE, para obtener los estimadores totales: '^Y' y '^V(^Y)'. Con APPEND estos nuevos valores se anexan a la derecha de las variables del archivo MC (14 variables en total); con la impresión de todas las variables (VAR=ALL).

La salida de CONG.LO es semejante al archivo temporal de salida del párrafo TRANSFORM, sólo que se anexan los valores de '^Y' y '^V(^Y)'.

(Se muestra sólo la tabla de resultados)

NDIS	Mi	yi1	yi2	yi3	yi4	M	mii	^Yi
5.0	14.0	2.0	3.0	2.0	5.0	150.0	4.0	42.0
8.0	12.0	4.0	8.0	9.0	7.0	150.0	4.0	84.0
10.0	24.0	1.0	1.0	6.0	4.0	150.0	4.0	72.0
11.0	6.0	4.0	3.0	3.0	6.0	150.0	4.0	24.0
12.0	8.0	1.0	4.0	8.0	7.0	150.0	4.0	40.0
15.0	13.0	2.0	5.0	5.0	4.0	150.0	4.0	52.0

pi	zi	AUX	^Y	^V(^Y)
0.93E-1	450.0	1.0	650.0	8500.0
0.80E-1	1050.0	1.0	650.0	8500.0
0.2	450.0	1.0	650.0	8500.0
0.40E-1	600.0	1.0	650.0	8500.0
0.53E-1	750.0	1.0	650.0	8500.0
0.87E-1	600.0	1.0	650.0	8500.0

3.3 Investigación 3.

Efecto del trasplante de tejido cerebral embrionario, en la recuperación funcional de ratas con Hemiplejía Experimental (disminución en la función de movimiento de la mitad del cuerpo).

(ESCALANTE y MÁRQUEZ, 1992)

En la actualidad existen evidencias de la recuperación de funciones perdidas por daño cerebral (Corral y Palacios 1983).

Para la recuperación de dichas funciones, en esta investigación se proponía el trasplante de tejido cerebral embrionario en el lugar dañado por los huéspedes. De aquí entonces; lo que se pretendía, era evaluar el efecto del trasplante de tejido cerebral embrionario en estado de suspensión en la recuperación del daño cerebral motor, considerando el modelo experimental de hemiplejía expresado en términos de recuperación inducida, en comparación con la de recuperación pasiva.

Se diseñó un experimento sobre una población de 24 ratas de tipo prospectivo, longitudinal y comparativo; longitudinal por el seguimiento en el tiempo y comparativo por la aplicación de cinco diferentes tratamientos a cinco grupos de ratas. (No obstante, más adelante, se hace notar, cómo este estudio puede prestarse, entre otros, para un estudio observacional. pág. 85).

Los investigadores asignaron aleatoriamente cinco diferentes tratamientos (control, lesión falsa, lesión verdadera, lesión verdadera más trasplante falso y lesión verdadera más trasplante verdadero) a dicha población, formando cinco grupos de ratas y fijando para uno de los grupos, un número de casos diferente al de los grupos restantes.

Así, los investigadores contemplaron que al grupo de menor número de casos, se le aplicara el tratamiento 1 (control), y a

los restantes, los demás tratamientos, provocando un experimento no balanceado. Esta situación así como el número de mediciones de las variables (123 para base, 169 para lesión y 212 para trasplante) y el número de días (45 para base, 60 para lesión y 75 para trasplante) en cada fase; fueron fijados por los investigadores, por razones fundamentales de su campo.

a) Objetivo general y presentación de la información:

Como objetivo general se propuso evaluar la potencialidad de los trasplantes de tejido cerebral embrionario en estado de disociación, para lo cual se emplea el modelo experimental de hemiplejía en ratas.

En el experimento se consideraron en estudio 24 ratas. Cada una estuvo bajo observación durante tres fases: Línea Base (LNB), Lesión (LES) y Trasplante (TRAS) y a cada rata se le asignó aleatoriamente sólo uno de los cinco tratamientos (Tabla 1).

TABLA 1: Número de rata por tratamiento aplicado.

Tratamiento	Número de rata
1 = Control	3, 9, 24, 4
2 = Lesión falsa *	23, 6, 8, 11, 15
3 = Lesión verdadera (lv)	10, 16, 20, 25, 21
4 = lv + trasplante falso **	13, 1, 12, 14, 18
5 = lv + trasplante verdadero	7, 17, 5, 2, 19

* Lesión falsa : Introducción de una cánula sin succión de tejido.

** Trasplante falso: Inyección de solución Ringer.

b) Variables para estudiar el efecto de los tratamientos:

Se aplicó una batería de pruebas para el momento de medición i en la fase K:

I. Análisis de la marcha de la rata o de corredor:

Largo del paso, Ancho del paso y ángulo del paso
 $LK(i)$, $AK(i)$, $DK(i)$

respectivamente.

II. Grid o rejilla: Tiempo requerido en soltarse de la rejilla

$GK(i)$

III. Nado. Tiempo de nado: $NK(i)$

IV. Fuerza en extremidades anteriores y posteriores

$EFK(i)$, $FPK(i)$

respectivamente.

V. Peso de la rata: $PK(i)$.

c) Variables codificadas en el estudio:

Las variables codificadas en el estudio son: NR (# de rata) primeramente, y enseguida un determinado número de mediciones de cada variable descrita anteriormente, en cada fase, es decir:

Para la fase K = LNB (línea base), las variables de los incisos del I al IV tienen 17 mediciones ($i=1,2,\dots,17$) para cada una de ellas, y 4 mediciones ($i=1,2,3,4$) para el inciso V (peso de la rata).

Para K=LES (lesión): $i=1,2,\dots,23$ para el inciso I a IV
 $i=1,2,\dots,8$ para peso; y

Para K=TRAS (trasplante):

$i=1,2,\dots,29$ para el inciso I a IV
 $i=1,2,\dots,9$ para peso.

O bien:

	K = LNB	K = LES	K = TRAS
$LK(i)$, $AK(i)$, $DK(i)$			
$GK(i)$			
$NK(i)$	$i=1,\dots,17$	$i=1,\dots,23$	$i=1,\dots,29$
$EFK(i)$, $FPK(i)$			
$PK(i)$	$i=1,\dots,4$	$i=1,\dots,8$	$i=1,\dots,9$
	$7(17)+4=123$	$7(23)+8=169$	$7(29)+9=212$
	total= 504 variables.		

Sean:

LB, AB, DB, GB, NB, EFB, FPB y PB las variables para la fase K=LNB,
LL, AL, DL, GL, NL, EFL, FPL y PL las variables para la fase K=LES y
LT, AT, DT, GT, NT, EFT, FPT y PT las variables para la fase K=TRAS;

cada una con su respectivo número de mediciones.

Se trata de las mismas variables, sólo que en diferente fase.

Así entonces, el registro formado es:

NR

LB1	LB2...LB17	AB1	AB2... AB17	DB1	DB2... DB17	GB1	GB2...GB17
NB1	NB2...NB17	EFB1	EFB2...EFB17	FPB1	FPB2...FPB17	<u>PB1</u>	<u>PB2... PB4</u>
LL1	LL2...LL23	AL1	AL2... AL23	DL1	DL2... DL23	GL1	GL2...GL23
NL1	NL2...NL23	EFL1	EFL2...EFL23	FPL1	FPL2...FPL23	<u>PL1</u>	<u>PL2... PL8</u>
LT1	LT2...LT29	AT1	AT2... AT29	DT1	DT2... DT29	GT1	GT2...GT29
NT1	NT2...NT29	EFT1	EFT2...EFT29	FPT1	FPT2...FPT29	<u>PT1</u>	<u>PT2... PT9</u>

Puede verificarse que el número total de variables por registro o rata es de 505: Para la línea base, 123; para lesión 169 y para trasplante 212, más una variable para número de rata (NR).

La información en estudio puede resumirse en la siguiente tabla:

TABLA 2

	día 0	día 45	día 105	día 180 (6 meses)
numero de rata	. FASE K=LNB (línea base) DURACION: 45 días VARS. MEDIDAS: LB, AB, DB, GB, <u>NB</u> EPB y FPB (i=1,17) PB (i=1,4)	. FASE K=LES (lesión) DURACION: 60 días VARS. MEDIDAS: LL, AL, DL, GL, <u>NL</u> , EFL y FPL (i=1,23) PL (i=1,8)	. FASE K=TRAS (trasplante) DURACION: 75 días VARS. MEDIDAS: LT, AT, DT, GT, <u>NT</u> , EFT y FPT (i=1,29) PT (i=1,9)	
3,9,24,4 (4 ratas)	Tratamiento 1: CONTROL: T1			
		noL		noT
23,6,8,11, 15 (5 ratas)		Tratamiento 2: LESIÓN FALSA: T2		
				noT
10,16,20, 25*, 21 (5 ratas)		Tratamiento 3: LESIÓN VERDADERA: T3		
				noT
13,1,12, 14,18 (5 ratas)			Tratamiento 4: T3+TRASPLANTE FALSO: T4	
		LV		TF
17,5,2,19 7 (5 ratas)			Tratamiento 5: T3+TRASPLANTE VERDADERO: T5	
		LV		TV

noL: no hubo lesión
noT: no hubo trasplante

LV: hubo lesión verdadera
TF: hubo trasplante falso
TV: hubo trasplante verdadero

* El número 24 para asignar a una rata, no existe.

3.3.1 Transposición y mezcla de archivos.

Transposición de archivos, con el programa SERIESTRAS.DM (uso del párrafo UNPACK), y enseguida mezcla horizontal de los archivos transpuestos con el programa SERIESFIN.DM (uso del párrafo JOIN).

En la práctica de la aplicación de la estadística es común que se solicite asesoría de análisis de datos para lo cual la mayoría de las veces el solicitante ya ha analizado parcialmente sus datos, por lo cual entonces, ya tiene un archivo.

En el transcurso de la asesoría se determina el análisis por realizar (tipo de estudio: transversal, longitudinal, y análisis estadístico), asimismo, se elige el software por utilizar; y si el análisis se puede realizar con el archivo de datos con el que ya se cuenta, entonces se dice que el archivo de datos es compatible con el análisis que se pretende realizar con el paquete elegido. Sin embargo, es usual que dicha compatibilidad no exista, por lo cual es necesario cambiar la estructura del archivo de datos, para hacerlo compatible con el análisis del paquete utilizado.

Dentro de este contexto, se puede percibir, la necesidad prioritaria de determinar el tipo de estudio, el análisis estadístico y el software.

Para el caso de este problema 3.3.1, se detallarán posibles análisis de tipo longitudinal y transversal, y dentro de estos, se especificará uno; para justificar el uso del software utilizado: transposición y mezcla, párrafos UNPACK y JOIN.

Como se ha mencionado, algunos estudios de investigación requieren a veces de un estudio longitudinal o transversal, o ambos. De manera general cuando se requiere analizar un estudio con ambos tipos de métodos es necesario tratar con el concepto de transposición de archivos de datos.

Para la investigación presente, pudiera ser posible determinar:

1. i) ¿Hay diferencia entre tratamientos?, ¿Qué variables permiten apreciar las diferencias entre tratamientos durante todo el período de estudio?. Respuestas que pueden ser determinadas a través de un Análisis Discriminante.
ii) Gráficos del comportamiento de las variables bajo estudio en el transcurso del tiempo. Gráficos de variables contra medición.
2. Si pensamos el estudio como longitudinal experimental sería de interés estudiar:
i) La variación de las variables bajo estudio por el efecto de tratamiento, fase y la interacción de éstas. Esto nos permitiría contestar si el efecto por tratamiento es el mismo, o si el efecto por fase es el mismo o si el efecto por tratamiento es distinto dependiendo de las fases en consideración (comparación de tratamientos en las fases). Dado que existe más de un caso que se estudia en el tiempo, este tipo de análisis es posible realizarlo con el de Análisis de varianza para mediciones repetidas.
3. Ahora, si el estudio lo pensáramos como longitudinal observacional, sería de interés:
i) Modelar las variables bajo estudio como Series de tiempo. Ahora bien, como realmente estos datos se generaron experimentalmente, entonces habría que considerar conjuntamente más de dos series de tiempo (ratas) en estudio; y no sólo una, como convencionalmente se hace.
Asimismo, sería necesario, incorporar una o ambas, de las siguientes funciones que se describen:
ii) Funciones de Transferencia. Explica una serie, no sólo en función de ella misma, sino también con otra serie. Ejemplo: Para la variable largo del paso, considerar como una serie explicativa el peso de la rata.
iii) Funciones de Intervención. Permite modelar las diferencias por fases en estudio (LV+TV) dentro de un tratamiento al considerar todas las fases en estudio (ejemplo, tratamiento=LV+TV). De esta forma se caracterizaría como una serie cada tratamiento, para luego comparar tal caracterización.

En el caso de requerir cambiar la estructura de cierto archivo de datos podría pensarse en volver a capturar la información ó, lo más conveniente, usar un software que cambie la naturaleza del archivo, máxime si existe un buen conocimiento de su manejo.

En este sentido, se describirá el estudio para analizar variables contra medición con el módulo 6D del BMDP, (programa JOVALAIN.6D) donde se puede apreciar una incompatibilidad respecto a la estructura del archivo; por lo que hay que 'poner a punto' el archivo de datos original de nombre TODOS, y llegar al TODOS2T, que es precisamente el archivo que necesita el módulo 6D.

En este caso, el módulo DM del BMDP, realizará esta tarea.

En el Programa JOVALAIN.6D, se aprecia que el módulo 6D requiere que las mediciones de las variables de las tres fases por graficar, estén expresadas en columna en el archivo de entrada TODOS2T.

Programa JOVALAIN.6D

```
/BMDP 6D
/PROBLEM TITLE IS 'GRAFICOS DE LAS VARIABLES'.
/COMMENT IS 'ETAPA : TODAS'.
/INPUT FILE = TODOS2T.VARIABLES = 11.FORMAT = FREE.
/VARIABLES NAMES = NR, SESIONES, LARGO, ANCHO, ANGULO, GRID, NADO,
                  FZADEL, FZATRA, PESO, GRUPO.
      GROUPING=NR.
/ GROUP CODES(NR)=3,9,24,4,    23,6,8,11,15,    10,16,20,25,21,
                  13,1,12,14,18,    17,5,2,19,7.
      NAMES(NR)=A,B, C,D,    E,F,G, H, I,    J, K, L, M, N,
                  O,P, Q, R, S,    T,W,X, Y,Z.
/PLOT YVAR = LARGO.
      XVAR = SESIONES.
      GROUPS = A,E.
      GROUPS = J,O.
      GROUPS = S,Z.
      GROUPS = A,B,C,D, E,F,G,H,I, J,K,L,M,N, O,P,Q,R,S, T,W,X,Y,Z
      MIN = 0.
      SIZE = 64, 18.
/PRINT LINESIZE = 80.
END/
FINISH/
```

Archivo TODOS2T:

*
NR SESNES LARGO ANCHO ANGULO GRID NADO FZADEL FZATRA PESO GRUPO

LK(1) AK(1) DK(1) GK(1) NK(1) EFK(1) FPK(1) PK(1)

		LB	AB	DB	GB	NB	EFB	FPB	PB
		(17 meds)							
3	1	lb1	ab1	db1	gb1	nb1	efb1	fpb1	pb1
		:	:	:	:	:	:	:	:
		lb17	ab17	db17	gb17	nb17	efb17	fpb17	pb17
		LL	AL	DL	GL	NL	EFL	FPL	PL
		(23 meds)							
		l11	a11	d11	g11	n11	ef11	fp11	p11
		:	:	:	:	:	:	:	:
		l123	a123	d123	g123	n123	ef123	fp123	p123
		LT	AT	DT	GT	NT	EFT	FPT	PT
		(29 meds)							
		lt1	at1	dt1	gt1	nt1	eft1	fpt1	pt1
		:	:	:	:	:	:	:	:
3	69	lt29	at29	dt29	gt29	nt29	eft29	fpt29	pt29
		LB	AB	DB	GB	NB	EFB	FPB	PB
		(17 meds)							
9	1	lb1	ab1	db1	gb1	nb1	efb1	fpb1	pb1
		:	:	:	:	:	:	:	:
		lb17	ab17	db17	gb17	nb17	efb17	fpb17	pb17
		LL	AL	DL	GL	NL	EFL	FPL	PL
		(23 meds)							
		l11	a11	d11	g11	n11	ef11	fp11	p11
		:	:	:	:	:	:	:	:
		l123	a123	d123	g123	n123	ef123	fp123	p123
		LT	AT	DT	GT	NT	EFT	FPT	PT
		(29 meds)							
		lt1	at1	dt1	gt1	nt1	eft1	fpt1	pt1
		:	:	:	:	:	:	:	:
9	69	lt29	at29	dt29	gt29	nt29	eft29	fpt29	pt29

(69x24)x11

* Nota: SESNES es un termino abreviado que corresponde a SESIONES
ver nota en la siguiente pagina.

Se observa que para cada rata se tiene el total de mediciones, representado en una columna, de una variable en las tres fases.

Otra forma de verlo, es que las mediciones de cada variable están transpuestas respecto a la forma original de captura de los casos o ratas (ver archivo TODOS, siguiente página), además de seguir otro orden, el de la continuidad de la misma variable, sin importar la fase, por ejemplo, LB, LL, LT.

Entonces el total de variables en una columna es de 69 mediciones en cada una de las ratas, y si se trata de 24 ratas, el número de renglones del archivo es 69x24; y el de columnas es 11 (NR, SESIONES, LARGO, ANCHO, ANGULO, GRID, NADO, FZADEL, FZATRA, PESO y GRUPO). La dimensión del archivo TODOS2T es, entonces, de (69x24)x11.

*Nota: Peso no es una variable que, en la realidad, contenga 69 mediciones (contiene 21), por lo que se agregan mediciones creadas para efecto del manejo del archivo.

La estructura anterior tuvo que ser formada a partir del archivo con que inicialmente se cuenta: TODOS, donde todas las mediciones de todas las variables de una rata específica se encuentran en un renglón:

Archivo TODOS

<u>NR</u>	<u>LB</u>	<u>AB</u>	<u>DB</u>	<u>...FPB</u>	<u>PB</u>	<u>LL</u>	<u>AL</u>	<u>DL</u>	<u>...FPL</u>	<u>PL</u>	<u>LT</u>	<u>AT</u>	<u>DT</u>	<u>...FPT</u>	<u>PT</u>	<u>PT10</u>	<u>GRUPO</u>
-----------	-----------	-----------	-----------	---------------	-----------	-----------	-----------	-----------	---------------	-----------	-----------	-----------	-----------	---------------	-----------	-------------	--------------

123 mediccs.

169 mediccs.

212 mediccs.

[3
9
25
:
19]



(24x507)

En un renglón se encuentra el total de mediciones en las tres fases de las variables LK(i), AK(i), DK(i), GK(i), NK(i), EFK(i), FPK(i) y PK(i), entonces, 123+169+212=504 variables; más NR, PT10 y GRUPO (éstas dos últimas son variables creadas), lo que da el total de 507. Por lo tanto, TODOS es un archivo de 24x507.

Se puede observar que del archivo TODOS se ha generado el TODOS2T. Este último expresa las mismas variables del archivo TODOS, sólo que transpuestas. Este proceso de cambiar los renglones a columnas (transposición), fue posible por la intervención del DM del BMDP.

Con el programa SERIESTRAS.DM, se aprecia cómo es posible realizar el procedimiento de la transposición. Enseguida se detalla este procedimiento, pero sólo para la variable FPT: fuerza en extremidades posteriores en la fase trasplante. Esto es, sólo se usará la parte sombreada del archivo TODOS.

Archivo de entrada de nombre TODOS, sólo con la variable FPT:

NR	FPT1	FPT2	FPT3	FPT29
1	fpt1,1	fpt1,2	fpt1,3 . . .	fpt1,29
2	fpt2,1	fpt2,2	fpt2,3 . . .	fpt2,29
3	fpt3,1	fpt3,2	fpt3,3 . . .	fpt3,29
			⋮	
25	fpt25,1	fpt25,2	fpt25,3 . . .	fpt25,29

(24x507)

Programa SERIESTRAS.DM

```

READ FILE= TODOS.
VARIABLES=NR,LBL1,for i=2 to 17. %LB|i,% for i=1 to 17. %AB|i,%
for i=1 to 17. %DB|i,%for i=1 to 17. %GB|i,%for i=1 to 17. %NB|i,%
for i=1 to 17. %EFB|i,%for i=1 to 17. %FPB|i,%
for i=1 to 4. %PB|i,%
for i=1 to 23. %LL|i,%for i=1 to 23. %AL|i,%
for i=1 to 23. %DL|i,%for i=1 to 23. %GL|i,%for i=1 to 23. %NL|i,%
for i=1 to 23. %EFL|i,%for i=1 to 23. %FPL|i,%
for i=1 to 8. %PL|i,%
for i=1 to 29. %LT|i,%for i=1 to 29. %AT|i,%
for i=1 to 29. %DT|i,%for i=1 to 29. %GT|i,%for i=1 to 29. %NT|i,%
for i=1 to 29. %EFT|i,%for i=1 to 29. %FPT|i,%
for i=1 to 9. %PT|i,%PT10,GRUPO.
FORMAT=FREE.CASE=24./
UNPACK VARIABLES=FPT1 to FPT29.#,% for i=1 to 17. %ab|i,%
#for i=1 to 17. %db|i,%for i=1 to 17. %gb|i,%for i=1 to 17. %nb|i,%
#for i=1 to 17. %efb|i,%for i=1 to 16. %fpb|i,%FPB17.
KEEP=NR. LEVELS=29.
#SEQUENCE=VAR,MED.
NEWVARIABLES=NR, MED, FPT. / #, AB, DB, GB, NB, EFB, FPB. /
PRINT VARIABLES=NR, MED, FPT. #AB, DB, GB, NB, EFB, FPB.
FIELD=2,2,6,2. / #,5.2,5.2,5.2,5.2,5.2./
#SAVE NEW.CODE=DM.SFILE=SERIES./
FINISH /

```

En el archivo TODOS, se suponen los registros ordenados en la variable NR.

Explicación del Programa SERIESTRAS.DM:

Se lee el archivo permanente de nombre TODOS.

En la instrucción VNames del programa SERIESTRAS.DM, se nombra a 507 variables: NR, más

123 mediciones de la fase 1,

(LB1, LB2, ..., LB17, AB1, AB2, ..., AB17, ... , FPB1, ..., FPB17, PB1, ..., PB4),

169 mediciones para la fase 2, (LL1, ..., LL23, . . . , PL1, ..., PL8),

212 mediciones para la fase 3, (LT1, ..., LT29, . . . , PT1, ..., PT9),

PT10 y GRUPO;

que serán leídas con formato libre sobre las 24 ratas (CASE=24).

El párrafo UNPACK, transpone sólo la variable FPT:

(UNPACK VARIABLES = FPT1 to FPT29.)

Hace caso omiso de las instrucciones posteriores al signo # .

NR, en la instrucción KEEP, será la variable donde se registrará en columna, el número de la rata tantas veces como mediciones haya para esa rata; en este caso, 29:

```
(fpt1,1  
fpt1,2,  
fpt1,3  
.  
.  
.  
fpt1,29).
```

LEVELS = . será el número de registros creados de archivo para cada rata.

Este número de registros se determina por el número de mediciones por transponer en cada rata.

Por lo tanto, cada registro de archivo por transponer, se convierte en tantos nuevos registros de archivo, como mediciones tiene la variable afectada por UNPACK, en este caso, FPT crea 29 nuevos registros para cada rata. (Ver archivo TODOS, siguiente página).

Con NEWNAMES = NR,MED,FPT , se estarán declarando los nombres para las columnas del archivo transpuesto. Tres nombres indican tres columnas.

En la primera se colocará el número de rata, en la segunda se enumerará la medición 1,2,...,29 , y en FPT , la tercera, irán las 29 mediciones ya transpuestas, para cada rata. (Ver archivo TODOS).

Con PRINT VARIABLES = NR,MED,FPT y su campo de escritura, (FIELD=) se imprimirá un archivo con tres columnas, una para cada de estas variables.

Archivo de salida TODOS:

rata	med.	FPT's
NR	MED	FPT
1	1	fpt1,1
1	2	fpt1,2
1	3	fpt1,3
	⋮	
1	29	fpt1,29
2	1	fpt2,1
2	2	fpt2,2
2	3	fpt2,3
	⋮	
2	29	fpt2,29
	⋮	
25	1	fpt25,1
25	2	fpt25,2
25	3	fpt25,3
	⋮	
25	29	fpt25,29

(24x29)x3

Entonces, SERIESTRAS.DM transpone una variable de una fase en las 24 ratas.

Una vez realizada la función para la variable FPT, se puede aplicar el programa a otras variables; es decir, aplicar el programa tantas veces como variables se requieran transponer.

De esta manera, pensemos en transponer las seis variables restantes de esa fase: LT, AT, DT, GT, NT y EFT.

Si se aplica un proceso SERIESTRAS.DM, para cada una de estas seis variables; se tendrán seis archivos de salida transpuestos, y finalmente, un total de siete archivos, contemplando el proceso para FPT.

Los seis archivos de salida transpuestos para las variables LT, AT, DT, GT, NT y EFT, de 29 mediciones de la fase trasplante en las 24 ratas, son:

archivo transpuesto para la var. LT	archivo transpuesto para la var. AT	archivo transpuesto para la var. EFT
1 1 lt1.1	1 1 at1.1	1 1 eft1.1
1 2 lt1.2	1 2 at1.2	1 2 eft1.2
.
1 29 lt1.29	1 29 at1.29	1 29 eft1.29
2 1 lt2.1	2 1 at2.1	2 1 eft2.1
2 2 lt2.2	2 2 at2.2	2 2 eft2.2
.
2 29 lt2.29	2 29 at2.29	2 29 eft2.29
.
25 1 lt25.1	25 1 at25.1	25 1 eft25.1
25 2 lt25.2	25 2 at25.2	25 2 eft25.2
.
25 29 lt25.29	25 29 at25.29	25 29 eft25.29

Los siete archivos pueden ser vistos en uno solo, para conjuntar las variables; lo que podrá lograrse al unir los siete archivos.

Para tal unión, cada archivo transpuesto funcionará como columna; realizándose entonces una unión de tipo horizontal, lo que equivale a agregación de variables en archivos.

Esta unión se logra con el Programa SERIESFIN.DM.

Programa SERIESFIN.DM:

```
# BMDP DM / *** MEZCLA VARIABLES POR RATA FASE LINEA TRAS***
FOR I= LT, AT, DT, GT, NT, EFT, FPT.
% READ SFILE = SER|I.
  FORMAT = FREE.
  VNAMES = NR, MED, V|I.
  FILE = SER|I. /
  SORT KEY = NR, MED. / %
JOIN FILES = SERLT, SERAT, SERDT, SERGT, SERNT, SEREFT, SERFPT.
  KEY = NR, MED.
  NEWFILE = ARCHSERI. /
PRINT FILE = ARCHSERI.      VARIABLES = ALL.
  FIELD = 2, 2, 9.2, 9.2, 9.2, 9.2, 9.2, 9.2. /
END /
FINISH /
```

Explicación del programa SERIESFIN.DM:

El programa SERIESFIN.DM lee con formato libre siete archivos permanentes de entrada de datos: SERLT, SERAT, ..., SERFPT, a través de las instrucciones

```
FOR I= LT, ..., FPT
% READ SFILE =SER|I.
FORMAT = FREE.
```

Se declaran los nombres de las variables NR, MED, VLT, VAT, VDT, VGT, VNT, VEFT y VFPT. Para la declaración de las siete últimas, se realiza con la instrucción: V|I .

Y para la denominación de los nombres de los siete temporales creados por el párrafo READ; se realiza con la instrucción:

```
FILE = SER|I
```

Estos temporales conservan el mismo nombre que los archivos permanentes: SERLT, SERAT, ..., SERFPT.

Antes de ejecutarse el párrafo JOIN, cada archivo temporal se ordena con respecto a las variables que hay en común en los siete archivos: NR y MED ; a través del párrafo SORT.

Tanto en el párrafo READ, SORT, como en las instrucciones que se encuentran entre los signos % , es extensiva la instrucción:

```
FOR I = LT, . . . , FPT
```

Finalmente junta los siete archivos temporales SERLT, SERAT,...,SERFPT respecto al número de rata y número de medición, con: JOIN FILES = SERLT , . . . , SERFPT.

KEY = NR, MED.

Al nuevo archivo de salida del párrafo JOIN se le nombra como ARCHSERI, el cual se imprime con el párrafo PRINT, en todas sus variables con los campos declarados por la instrucción FIELD.

Se forman así nueve columnas para el archivo ARCHSERI; las dos primeras corresponden a las variables en común que existen en los siete archivos, y las siete restantes a las mediciones de las variables ya transpuestas.

Archivo temporal ARCHSERI:

rata		med						
		LT	AT	DT	GT	NT	EFT	FPT
1	1	lt1,1	at1,1	dt1,1	gt1,1	nt1,1	eft1,1	fpt1,1
1	2	lt1,2	at1,2	dt1,2	gt1,2	nt1,2	eft1,2	fpt1,2
1	3	lt1,3	at1,3	dt1,3	gt1,3	nt1,3	eft1,3	fpt1,3
	:							
1	29	lt1,29	at1,29	dt1,29	gt1,29	nt1,29	eft1,29	fpt1,29
2	1	lt2,1	at2,1	dt2,1	gt2,1	nt2,1	eft2,1	fpt2,1
2	2	lt2,2	at2,2	dt2,2	gt2,2	nt2,2	eft2,2	fpt2,2
2	3	lt2,3	at2,3	dt2,3	gt2,3	nt2,3	eft2,3	fpt2,3
	:							
2	29	lt2,29	at2,29	dt2,29	gt2,29	nt2,29	eft2,29	fpt2,29
	:							
25	1	lt25,1	at25,1	dt25,1	gt25,1	nt25,1	eft25,1	fpt25,1
25	2	lt25,2	at25,2	dt25,2	gt25,2	nt25,2	eft25,2	fpt25,2
	:							
25	29	lt25,29	at25,29	dt25,29	gt25,29	nt25,29	eft25,29	fpt25,29

Nota: En este procedimiento de las variables de la fase trasplante, no se incluye la variable PT, pero puede tomarse en cuenta considerando que contiene 29 mediciones en esta fase, 23 en la de lesión y 17 en línea base.

Como puede observarse, SERIESTRAS.DM es un programa para ir obteniendo parcialmente la transposición, y SERIESFIN.DM conformará la matriz transpuesta final.

Los siete procesos SERIESTRAS.DM y el proceso SERIESFIN.DM pueden ser aplicados a la fase línea base, dando como resultado un archivo que pudiera ser llamado ARCHSERB.

De la misma manera, siete procesos SERIESTRAS.DM y un proceso SERIESFIN.DM para la fase lesión, pudieran crear un archivo llamado ARCHSERL.

Finalmente se puede proponer que se realice un programa que una por renglón estos tres últimos archivos: ARCHSERI, ARCHSERB y ARCHSERL, intercalando los registros (párrafo MERGE, es decir, los registros de un archivo con los de otro(s), mezcla vertical, o bien, el final de un archivo con el inicio de otro); para conformar el archivo TODOS2T, pero ordenado en la variable NR; y con exclusión de la variable PK.

Los siguientes esquemas ayudan a entender, cómo se forma la estructura del archivo TODOS2T.

Archivo ARCHSERB (base)

LB AB DB GB NB EFB FPB

1	1
...	...
1	17
2	1
...	...
2	17
...	...
25	1
...	...
25	17

Archivo ARCHSERL (lesión)

LL AL DL GL NL EFL FPL

1	1
...	...
1	23
2	1
...	...
2	23
...	...
25	1
...	...
25	23

Archivo ARCHSERI (trasplante)

LT AT DT GT NT EFT FPT

1	1
...	...
1	29
2	1
...	...
2	29
...	...
25	1
...	...
25	29

Como se mencionó; se aplica el párrafo MERGE, intercalándose los registros de los tres archivos anteriores, y así resulta el archivo TODOS2T, sólo que el que se refiere ahora, está ordenado en la variable NR:

Archivo TODOS2T (ordenado en NR)

1	1	} 69 mediciones (SESIONES)
.	.	
.	.	
1	17	
1	1	
.	.	
.	.	
1	23	
1	1	
.	.	
1	29	
.	.	
.	.	
25	1	
.	.	
.	.	
25	17	
25	1	
.	.	
.	.	
25	23	
25	1	
.	.	
.	.	
25	29	

CAPÍTULO 4.

CONCLUSIONES

En el módulo DM, como se ha podido observar, se tiene una gran flexibilidad en la preparación de archivos de datos, previo al análisis estadístico. En particular, en el trabajo se ha enfatizado la:

- 1 . Actualización de archivos temporales
- 2 . Conjunción de variables de información muestral o poblacional
- 3 . Conjunción de casos-muestras de una población
- 4 . Elaboración de gráficas o mapa de datos faltantes
- 5 . Generación de registros agregados o desagregados
- 6 . Realización de operaciones por renglón o por columna
- 7 . Transposición de los archivos para un estudio específico: longitudinal o transversal
- 8 . Lectura de archivos no rectangulares

Desarrollo de los puntos anteriores.

1. Este apartado es de gran importancia puesto que la creación de información temporal donde ya se han ejecutado operaciones en un archivo de datos, no se pierde; esto es, puede ser accesada en el momento y lugar que se requiera. A lo largo de este trabajo, puede observarse esta característica; desde la referencia en el desarrollo del párrafo AGGREGATE, (página 24), como en la estructura de los programas utilizados en los capítulos tratados.

Para ello utiliza 20 párrafos de instrucciones, los cuales pueden estructurarse en una gran variedad de formas, permitiendo la separación de partes en problemas complejos, para luego ensamblarlas paso a paso de una manera lógica.

La creación de los archivos temporales permite lograr este objetivo.

2. Es frecuente encontrar que los datos de una investigación no se recolectan en el mismo tiempo o han sido formados de diferentes fuentes o ambas cosas.

Por ejemplo, a un médico que lleva el registro de la información de sus pacientes cada vez que se presentan a consulta o requerir el historial médico de alguna otra fuente, le gustaría manejar esta información conjuntamente; lo cual se logra con la intervención del párrafo JOIN.

3. Situándonos en el mismo ejemplo del médico; si éste requiere agregar a su información nuevos pacientes, necesita juntar estos nuevos casos a los ya existentes; lo que logra con el uso del párrafo MERGE, cuya función es precisamente adicionar renglones a un archivo.

4. Gráficos o mapas son de gran utilidad para conocer la estructura de archivos de información de gran magnitud.

Si el médico, del ejemplo anterior, requiere aplicar diferentes tratamientos a cada uno de sus pacientes a lo largo de varias semanas, registrar tal información y visualizar toda su información en una forma descriptiva, lograría su objetivo con el uso del párrafo MAP.

5. Si ahora al médico le interesara evaluar algunas características médicas para obtener un reporte en forma sintetizada para cada paciente, sería de gran utilidad la aplicación del párrafo AGGREGATE.

6. El párrafo TRANSFORM, permite realizar operaciones sobre los renglones de un archivo, obteniendo entre otras, media, desviación estándar.

Por ejemplo, estima parámetros poblacionales en diversos esquemas de muestreo al usarse conjuntamente con las funciones estadísticas de DM.

Un párrafo TRANSFORM no opera con otro registro hasta que termina con uno, mientras que AGGREGATE opera sobre las variables de los registros que necesite, y lo interesante es que puede usarse después la información del TRANSFORM para el párrafo AGGREGATE.

Un ejemplo claro se dio en los diseños de muestreo cuando se obtuvieron los estimadores en cada unidad primaria (Y_i), para obtener posteriormente los estimadores poblacionales (Programa CONG.LO).

7. La solución a los problemas de estudios longitudinales y transversales puede ser resuelta con la aplicación del módulo DM. El hablar de estudios longitudinales y transversales es lo mismo que pensar en archivos de datos que pueden ser "volteados" o transpuestos. Estas funciones las proporcionan los párrafos PACK y UNPACK.

Aunque estos párrafos no siguen la forma convencional de transposición conocida, es posible llegar a ella con la aplicación de otros párrafos de DM. Este caso es claro en la Investigación 3.

8. Existen formas de transposición donde los archivos de entrada y de salida coinciden en ser rectangulares.

Una situación donde no ocurre esto, es precisamente la que se menciona para la obtención del archivo TODOS2T (página 87); donde resultaría un archivo no rectangular de no "rellenar" la variable PK; es el caso donde un archivo rectangular produce uno no rectangular, al ser transpuesto.

La variable peso, PK, es la única que tiene diferente número de mediciones respecto al grupo de las siete restantes en cada fase: por ejemplo, en la fase línea base; las variables LB,...,FPB tienen 17 mediciones cada una de ellas, mientras que PB tiene cuatro. Esta situación no crea problema en un archivo como TODOS, (página 88) pues el total de mediciones de todas las variables, se contempla en un renglón.

En cambio, en un archivo como TODOS2T, donde se conjunta información transpuesta de las tres fases; se requiere que todas las variables dentro de cada fase tengan el mismo número de mediciones, debido a que al transponer las tres fases conjuntamente, se forma un total de renglones igual para las variables.

El archivo transpuesto general contaría con registros diferentes, pues algunos renglones incluirían diferentes variables; en este caso algunos incluirían la variable PK y otros no.

Esto ocasiona un archivo no rectangular, puesto que los registros son de diferente tipo (algunos renglones tienen variables que otros renglones no tienen).

En la investigación se solucionó este problema, "rellenando" los lugares de PK en el archivo TODOS2T hasta completar 69 renglones en su columna.

Lo importante de estos casos radica en tener una solución cuando se presenten los archivos no rectangulares (puede tratarse el párrafo EXTRACT, por ejemplo) ya que DM sólo salvará archivos rectangulares.

Casos como estos pueden proponerse como continuidad para ser resueltos por medio del DATA MANAGER; lo cual permite reafirmar la importancia de la potencia del módulo DM.

Actualmente algunos paquetes estadísticos de prestigio, incorporan un manejador de datos. Esto ha sido de dominio público, cuando por ejemplo, en el Symposium 18avo. de Colorado U.S.A. (1986), se convocó a vendedores de 'software' para que presentaran la solución a cuatro problemas de manejo de archivos de datos. Los paquetes que presentaron solución en ASA 1986, fueron: PRODAS, P-STAT, SAS, SPSS, SYSTAT y BMDP. Este último presenta su solución en el reporte técnico #84.

La revisión de estos documentos, permite apreciar la necesidad que se tiene en la aplicación de la estadística de preparar los datos cuando uno pretende analizarlos.

Cabe destacar como último punto lo referente a la elaboración de tablas o cuadros estadísticos por medio de la ayuda del BMDP:

Cualquier investigación posibilita el desarrollo de trabajos de investigación alrededor de la construcción de tablas o cuadros estadísticos; puesto que se trata del lugar donde pueden ser presentados los resultados generados de una investigación.

Estos reportes constituyen en la práctica los resultados finales de una investigación. A este respecto, el módulo DM del BMDP permite obtener tablas y frecuencias, y módulos como el 4F obtiene tablas de contingencia. (El contenido de las celdas de una tabla de contingencia es una variable cuantitativa).¹

Por lo anterior es que, en la actualidad, puede pensarse en un paquete estadístico ideal, como aquél que pueda preparar los datos, procesarlos estadísticamente y reporte tablas o cuadros estadísticos de resultados.

¹ Ver referencia de Elaboración de tablas o cuadros estadísticos.

B I B L I O G R A F Í A

BMDP Statistical Software Manual. The Data Manager. Vol. 3. University of California Press. Berkeley. Los Angeles. Oxford. Copyright 1992 by BMDP Statistical Software, Inc.

COCHRAN, W., (1972). *Técnicas de Muestreo.* Compañía Editorial Continental, S.A. 3a. impresión. Impreso en México.

CUEVAS, C.S. (1990). *Investigación sobre Salud Mental y Proceso Migratorio.* Resultados y aspectos técnicos. Anales de la V reunión de Investigación y Enseñanza. Instituto Mexicano de Psiquiatría.

ENGELMAN, L., HILL, M. (January 1987). *Using the BMDP Data Manager: Examples of Four File Problems. Technical Report #84.* BMDP Statistical Software, Inc. Los Angeles, California 90025.

ENGELMAN, L., HILL, M., GUTHRIE, D., KRUPP, J., MEYER, J. (1990). *BMDP Data Manager Manual (To accompany the 1990 software release).* BMDP Statistical Software, Inc. Los Angeles.

ESCALANTE, M. Ma. del C., MÁRQUEZ, R. R. (1992). *Recuperación Funcional en Ratas con Hemiplejía Experimental.* (Investigación Psicobiológica básica). Tesis de Licenciatura. México. Facultad de Ciencias. U.N.A.M.

KISH, L. (1975). *Muestreo de Encuestas.* Editorial Trillas. México. D.F.

KOJI, A. (1987). *Desarrollo de un conjunto de Rutinas para microcomputadoras para la resolución de problemas básicos de muestreo*. Tesina de Postgrado. U.A.C.P. y P. del C.C.H. I.I.M.A.S. U.N.A.M.

MÉNDEZ, R. I., NAMIHIRA, G. D., MORENO, A. L., SOSA DE, M. C., CAÑEDO, D. L., SHABOT, A. E. *El Protocolo de la Investigación*. Edición Preliminar, México. D.F.

PABLOS, H. J.L., DEL RÍO J. I. (1981). *Un lenguaje para descripción de tablas*. Serie metodología 2. Centro Nacional de Información y Estadísticas del trabajo (CENIET). Secretaría del Trabajo y Previsión Social.

TREJO, V. B. (1985). *Introducción a Técnicas de Muestreo: Ejercicios de Aplicación de Temas de un curso básico*. Comunicación Técnica No. 29. Serie verde: Notas. I.I.M.A.S. U.N.A.M