CENTRO LATINOAMERICANO DE DEMOGRAFIA

Banco de Datos

# PROGRAMAS DE COMPUTACION USADOS EN EL CELADE

1

CELADE

AGOSTO 1972

CENTRO LATINOAMERICANO DE DEMOGRAFIA
Banco de Datos

# PROGRAMAS DE COMPUTACION

# USADOS EN CELADE

Serie J
N° 1
500

Agosto 1972

# PRESENTACION

Con esta nueva serie de publicaciones del Banco de Datos, sobre aspectos metodológicos de la computación electrónica, se espera cumplir con los siguientes objetivos:

1. Informar a los usuarios del Banco de Datos qué programas de computación hay disponibles, a fin de satisfacer demandas de tabulaciones especiales.

2. Ofrecer información en español que pueda servir a personas que trabajan en procesamiento de datos de encuestas o censos, permitiéndoles el uso eficiente de algunos programas o conjuntos de programas ya confeccionados y la obtención, en forma rápida de los resultados requeridos.

3. Contar en CELADE con el material didáctico necesario para ser utilizado en seminarios o cursos relativos a elaboración o tratamiento de datos.

En este primer volúmen se ha pretendido alcanzar sólo el primero de los objetivos, para algunos de los programas disponibles: "Statistical Package for the Social Sciences" (SPSS), "Organized Set of Integrated Routines for Investigation with Statistics" (OSIRIS/40), "MINI-TAB" y "A Program Language" (APL).

En las próximas publicaciones de la serie se entregará información más detallada sobre cada uno de estos programas u otros sistemas con que cuente CELADE.

# I N D I C E

# I. DESCRIPCION DE LOS PROGRAMAS DISPONIBLES EN APL

## ("A Program Language")

### 1. Introducción

Existen sistemas que permiten el uso simultáneo de un computador por una gran cantidad de usuarios. Esto se hace a través de "terminales", que son máquinas de escribir convenientemente acondicionadas para comunicarse en forma directa con el computador central, utilizando un cierto "lenguaje de terminales".

La utilización simultánea de un computador por muchos terminales, o sea, por muchos usuarios diferentes, es posible gracias a los "sistemas de tiempo compartido" que aprovechan al máximo el tiempo de uso de la unidad central de proceso del computador (CPU), atendiendo a cada uno de los distintos terminales por fracciones de segundo. Esto da la impresión de que cada terminal está utilizando el computador en forma privada, y en la práctica es como si así fuera.

En Chile, la IBM ha instalado, con la ayuda de un computador IBM 360/40 H, el sistema APL que funciona con un lenguaje propio (APL/360) y con terminales tipo 2741.

El sistema trabaja en modo conversacional: el usuario transmite una cierta orden, el computador la ejecuta "inmediatamente" y envía la respuesta. Estas órdenes pueden ser instrucciones individuales o grupos de instrucciones (programas) que deben ser definidos previamente como tales. El sistema trabaja en dos modos:

a) Modo de ejecución, en el cual el computador envía inmediatamente la respuesta a cada instrucción; y

b) Modo de definición, en que el usuario conforma un programa, agrupando instrucciones que serán procesadas en conjunto al "ordenarse" su ejecución. Cada uno de estos programas tiene internamente una característica que los individualiza. En este caso la respuesta se entrega una vez terminada la ejecución de todas las instrucciones del programa. Eso hace posible plantear complicados cálculos matemáticos para resolver gran cantidad de problemas.

Este sistema permite resolver problemas matemático-lógicos en un computador que prácticamente se tiene disponible como si perteneciera a la misma oficina.

Como lo normal es que estos programas se utilicen más de una vez, se cuenta con "áreas de memoria" a disposición de cada usuario, en donde se pueden guardar por tiempo indefinido los programas, "bibliotecas privadas" y los datos y resultados que sean de interés.

Cualesquiera de estos valores y programas pueden ser modificados a voluntad, permitiendo realizar análisis de problemas con un rendimiento mucho mayor que el que se logra empleando programas en tarjetas perforadas en un computador, pues se obtienen resultados con mayor rapidez mediante la reducción del tiempo de respuesta. Además, se pueden conocer los resultados intermedios al proceso, lo que permite hacer un análisis mucho más dinámico que de la manera tradicional. Esto es útil sobre todo cuando se están probando distintas hipótesis, a fin de llegar a la metodología más apropiada para enfrentar un problema.

Otra de las ventajas importantes que tiene este sistema es que una vez elaborado el programa, éste puede ser utilizado por personas que no sepan cómo lo ejecuta el computador, no conozcan el lenguaje empleado ni el programa que se usa. Esto es válido aún en programas que necesitan diferentes tipos de datos y en distintos momentos del proceso, pues el programa se puede hacer de modo que él mismo solicite los datos que vaya necesitando, y que explique en qué orden debe entregárselos quien esté frente al terminal. Esta facilidad de uso de programas ya elaborados hace muy atractivo el empleo de sistemas conversacionales como el APL.

En el trabajo con el terminal instalado en CELADE se ha podido notar la importancia de estas características, al programar y utilizar este sistema en la resolución de una gran y variada cantidad de problemas.

En las páginas siguientes se presenta un breve resúmen de los programas que se encuentran en "bibliotecas privadas" de CELADE y en las "bibliotecas públicas" de APL. Estas últimas son áreas de memoria que contienen programas elaborados por los programadores de IBM y que se encuentran a

disposición de cualquier usuario. Los programas que en ellas se encuentran abarcan prácticamente todos los tópicos del cálculo numérico, además de programas especialmente preparados para la enseñanza del APL/360 a los usuarios.

En general, los programas que se han hecho en CELADE son de índole demográfica, más otros de cálculo general que se utilizan regularmente en diferentes investigaciones. La parte teórica de los primeros ha sido desarrollada por investigadores y becarios del Centro, y su programación estuvo a cargo de programadores del Servicio de Computación.

Los programas referidos a continuación han sido clasificados en:

- Programas disponibles en la "biblioteca privada" de CELADE.

a) De índole demográfica;

b) De cálculo general

- Programas disponibles en la "biblioteca pública" de APL, de uso más frecuente en CELADE.[1/]

---

1/ Para mayor información, remitirse a las siguientes publicaciones:

- APL/360 Users Manual. H20-0683, IBM
- APL/360 Primer. H20-0689, IBM
- APL/360 Reference Manual, Sandra Pakin
- Apuntes de APL/360. Santiago Vásquez, Centro de Computación, Universidad de Chile. Publicación 70-9
- Bibliotecas Públicas para APL/360. Centro de Computación. Universidad de Chile. Publicación 70-8
- Guía de APL, Lenguaje de Terminales. Jacobo Gordon Universidad de la Plata. Argentina.

## 2. Programas disponibles en la "biblioteca privada" de CELADE

### a) De índole demográfica

"GREVILLE": Calcula cuatro valores interpivotales, aplicando multiplicadores de Greville. Los datos para este programa se entregan a través de un vector de cualquier cantidad de elementos, a los cuales se aplican los multiplicadores extremos, semiextremos y centrales de Greville. Los valores que así se obtienen son los de una curva de "suavidad óptima" trazada por los valores datos.

"SPRAGUE": Desglosa los datos de un total o subtotal en 5 partes (por ejemplo, obtiene edades individuales en base a grupos quinquenales) mediante la aplicación de los multiplicadores extremos, semiextremos y centrales de Sprague. El vector de datos puede tener tantos elementos como se desee. Los valores desglosados presentan también una curva suave.

"PROY": Permite realizar proyecciones de población por edades individuales en base a una población inicial (vector) y a sus relaciones de supervivencia correspondientes (matriz). Permite calcular también, opcionalmente, nacimientos en base a tasas específicas de fecundidad y otros datos afines, que pide oportunamente. Entrega finalmente los grupos quinquenales de población que resultan de las proyecciones individuales.

"TASAS": Calcula tasas, suponiendo una variación exponencial entre los datos. Este programa va pidiendo al operador, uno por uno los datos que se utilizarán, y posteriormente, una vez ingresados todos ellos, consulta sobre las tasas que interesan. Entrega, además, las fechas-datos, transformando meses y días en fracciones de año.

"TABLAMX": Conforma una tabla abreviada de mortalidad en base a tasas específicas de mortalidad por el método de Reed y Merrel. Calcula, además, esperanzas de vida y relaciones de supervivencia que se desprenden de esta tabla.

"TABLAQX": Realiza los mismos cálculos de la TABLAMX, usando el vector de probabilidades de muerte y empleando relaciones adecuadas para este tipo de datos.

"RELACSUPERV": Calcula relaciones de supervivencia entre poblaciones.

"VIUDAS 1" y "VIUDAS 2": Realizan una serie de análisis de mortalidad en base a nupcialidad, considerando la viudez como variable determinante en el proceso. Este es un programa de investigación del Centro.

"PXQUINQT": Calcula relaciones de supervivencia quinquenales en base a relaciones de supervivencia por edad detallada, que se ingresan como una matriz de dos dimensiones, de cualquier tamaño.

"PXIND": Calcula relaciones de supervivencia por edad detallada y por años calendario, en base a relaciones de supervivencia para grupos quinquenales y para años terminados en 0 y 5.

"GRQ": Agrupa una matriz de población por edades individuales, en matriz de grupos quinquenales.

"RELSUPACUM": Calcula relaciones de supervivencia acumuladas, en base a población inicial y final.

b) De cálculo general

"SRV": Realiza correlaciones simples entre dos vectores de observaciones y entrega una matriz con gran cantidad de información estadística sobre los datos. Este programa, al que se le ha agregado instrucciones a fin de obtener mayor información de tipo estadístico, está basado en el programa "SR" de la biblioteca pública de APL.

"LOGITO": Calcula la función "Logito" de los datos, los cuales pueden ser ingresados ya sea como un valor individual, como vector, o como matriz de cualquier dimensión.

"ANTILOGITO": Es la función inversa de la función LOGITO, y trabaja en forma análoga a ella.

"MGMAT": Calcula la matriz media geométrica de la matriz de datos.

"INTERPOLA 4": Interpola geométricamente cuatro valores entre el vector de datos iniciales y el vector de datos finales. Entrega una matriz con los valores interpolados.

"DESACUMULA": Obtiene vector de valores desacumulados desde el vector de datos (valores acumulados).

"NEWTON": Realiza interpolaciones entre valores-datos mediante fórmula de Newton.

## 3. Programas disponibles en la "biblioteca pública" de APL

Se mencionarán sólo aquellas funciones que se han usado con mayor frecuencia, o que tienen alguna aplicación para la investigación social.

"DSTAT": Estadística descriptiva. Para un vector de observaciones calcula y lista tamaño de la muestra, valores máximo y mínimo observados, rango, media, varianza, desviación estándar, desviación media, mediana y moda.

"FR": Tabla de frecuencia de una entrada.

"FR2": Tabla de frecuencia de dos entradas.

"CTAB": Tabla de contingencia de dos entradas. Calcula Chi cuadrado y los grados de libertad para una Tabla de Contingencia de dos entradas.

"MREG": Es un conjunto de funciones que permiten calcular regresión lineal múltiple. Entrega, además, medias, varianza, desviaciones estándar, correlaciones simples y correlaciones parciales, residuos y test de residuos.

"PLOT": Gratificación de funciones. La función gratificará simultáneamente una o más funciones, ajustando los valores aproximados a las dimensiones de escala especificados por el usuario.

"INV": Inversión de matrices. Calcula la inversa de una matriz cuadrada, empleando para ello el método de Gauss Jordan con pivote de inversión de matrices (se supone la matriz de entrada no singular).

# III. DESCRIPCION DE LOS PROGRAMAS SPSS

## ("Statistical Package for the Social Sciences")

### 1. Introducción

El SPSS, desarrollado en la Universidad de Chicago, es un sistema integrado de programas para el análisis de datos en ciencias sociales. El sistema ha sido diseñado para proveer al científico social de un paquete amplio y unificado, capaz de realizar diferentes tipos de análisis de datos de manera simple y conveniente. El SPSS posee una gran flexibilidad, en cuanto al formato de los datos. Provee al usuario de un amplio conjunto de procedimientos para transformaciones de datos y manipuleo de archivos y ofrece al investigador una gran cantidad de rutinas estadísticas usadas comúnmente en las ciencias sociales.

Además de los indicadores estadísticos usuales, distribución de frecuencias simple y tabulaciones cruzadas, el SPSS contiene procedimientos para correlación simple, correlación parcial, regresión múltiple, análisis factorial y "Guttman scaling". La facilidad de su manejo permite utilizar el programa para modificar permanentemente un archivo de datos conjuntamente a cualquiera de los procedimientos estadísticos. Estas facilidades permiten al usuario generar transformaciones de variables, recodificar variables, muestrear, seleccionar o hacer "pesar" casos específicos y para agregar o alterar datos al archivo que entrega la información. El SPSS hace posible realizar análisis mediante el uso de instrucciones de control en lenguaje inglés natural y no requiere experiencia en programación, de parte del usuario.

### 2. Propósitos de un paquete de programas estadísticos

Los computadores son extremadamente útiles en el procesamiento de gran cantidad de datos. De hecho, la necesidad de elaborar informaciones en gran escala ha influído directamente en el desarrollo de estos equipos. Las tareas de clasificación, ordenamiento (sort), almacenamiento y recuperación de datos que han sido presentados al computador codificados en

una forma conveniente, llamadas procesamiento o elaboración de datos, constituyen uno de los más importantes usos de los computadores en el presente.

Debido a su capacidad para llevar a cabo operaciones aritméticas a alta velocidad, los computadores son también muy usados para llevar a cabo cálculos matemáticos largos. Cuando tales cálculos son ejecutados con el propósito de analizar datos, se usa con frecuencia el término "análisis de datos". El análisis de datos combina la elaboración con el uso de procedimientos matemáticos o estadísticos.

Comúnmente, en jerga computacional, se hace diferencia entre aplicaciones de tipo comercial y de tipo científico. Las aplicaciones comerciales, generalmente, requieren gran cantidad de datos de entrada y salida y una pequeña cantidad de cálculos, mientras que las aplicaciones de tipo científico abarcan relativamente poca cantidad de datos de entrada y salida y gran cantidad de cálculos. Si uno acepta esta distinción, entonces el análisis de los datos, particularmente en el contexto de las ciencias sociales, quedaría entre estos dos extremos.

El análisis de los datos de las ciencias sociales frecuentemente envuelve la aplicación rutinaria de cierto número de operaciones. Cuando se usa un computador, es necesario detallar la secuencia exacta de los pasos a seguir en cada etapa del procedimiento. A la secuencia de pasos se conoce como un programa. Una vez que el programa está listo, puede aplicarse a varios conjuntos diferentes de datos, mediante ajustes menores, los cuales se pueden hacer a través de tarjetas de control de programas. Los centros de computación mantienen bibliotecas de programas ya preparados que el usuario puede utilizar para llevar a cabo ciertos procedimientos estándares.

Si el usuario está muy familiarizado con el análisis de datos habrá notado la repetición de una variedad de procedimientos. Habiendo procesado sus datos con un programa, puede desear usar los datos de salida de ese programa como entrada de otro. Puede requerir una gran cadena de tales trabajos. En ese caso, es importante que la salida de un programa sea compatible con la entrada de otro. Si se están usando varios programas, y ellos operan de maneras muy diferentes, se deberá dominar en detalle muchos programas, con lo que aumentan las posibilidades de error u omisión.

Un sistema (como el SPSS) es un conjunto de programas que ejecutan ciertos procedimientos relacionados y que comparten algunas convenciones referentes a la manera como el sistema maneja los datos. Si el sistema está bien diseñado, permitirá al usuario ejecutar una secuencia de tareas con un mínimo de intervención manual. El SPSS es un conjunto de programas relacionados que tiene como finalidad el manejo y análisis estadístico de muchos tipos de datos, con particular énfasis en el campo de las ciencias sociales. En adelante, nos referiremos a los programas del sistema como sub-programas. Una vez que el usuario ha introducido sus datos nuevos en el sistema, puede instruir al computador para que con los sub-programas del SPSS lleve a cabo una variedad de trabajos relacionados, en la secuencia que sea conveniente, según las circunstancias. No es preciso que el usuario vuelva a introducir los datos cada vez, puesto que el sistema los guardará y los usará apropiadamente cuando se los requiera. Los datos del sistema se pueden recuperar y utilizar en cualquier nuevo programa.

Se ha intentado incluir en el sistema SPSS los procedimientos más comúnmente usados en el análisis de datos en las ciencias sociales. El sistema SPSS puede ser agrandado para incluir procedimientos que no han sido previstos y que algún usuario considere conveniente agregar.

El SPSS provee un conjunto de convenciones comunes, válidas para diferentes programas, y que constituye un lenguaje simplificado correspondiente al que un científico social podría usar para describir los procedimientos que él desee utilizar.

En las páginas siguientes damos una breve descripción de los principales programas de SPSS. 2/

---

2/ S.P.S.S.
Statistical Package for the Social Sciences
Norman H. Nie, Dale H. Bent, y C. Hadlai Hull
Mc. Graw Hill, 1970

## 3. Procedimientos estadísticos del SPSS

El propósito de cualquier proceso de análisis de datos es condensar la información en un conjunto de datos ordenado sistemáticamente; de manera que facilite su interpretación. En el fondo, trata de obtener modelos de relaciones entre conjuntos de variables; o sea, es un medio para construir y experimentar teorías sociales empíricas.

El SPSS contiene la mayoría de los procedimientos más comunmente empleados en investigación social.

Quizás el mejor medio de catalogar los procedimientos estadísticos disponibles en el SPSS sea hacerlo de acuerdo a la función que realizan en el proceso de análisis de datos. Por lo tanto, se los presentarán de acuerdo a su nivel de complejidad y sofisticación.

### a) Distribuciones de frecuencia de una entrada, medidas de tendencia central y dispersión.

En la mayoría de las investigaciones sociales, el primer trabajo de análisis de los datos es el examen de las características de distribución de cada una de las variables que están bajo investigación.

El SPSS contiene tres procedimientos estadísticos con este fin: CONDESCRIPTIVE, que está diseñado para ser utilizado con variables con muchos intervalos de clase, lo cual supone un gran número de valores; y los dos subprogramas CODEBOOK y MARGINALS que se han diseñado para ser usados con variables que suponen sólo un número limitado de valores.

Estos tres subprogramas (CONDESCRIPTIVE, CODEBOOK y MARGINALS) pueden producir, opcionalmente, medidas estadísticas como el promedio, la moda, valores mínimos y máximos, desviación estándar y rango, según el interés del usuario. El CODEBOOK puede también entregar histogramas de distribución de frecuencias. En resúmen, CODEBOOK provee al investigador de información que compiló, inicialmente, a fin de determinar con qué tipos de datos se cuenta. (Reproduce cualquier

carácter que se haya utilizado en la codificación de los datos, distinguiendo entre blancos y ceros).

b) Tablas de dos o más entradas

Después que el investigador ha averiguado las características de cada una de las variables, puede continuar investigando grupos de relaciones. En el SPSS se han seleccionado varios procedimientos para examinar relaciones que el investigador podrá utilizar, según sean las características de las variables estudiadas y los propósitos que persiga. Podrá escoger análisis de correlaciones o alguna forma de despliegue de tablas de las que se indicarán más adelante.

El SPSS comprende dos procedimientos: CROSSTABS y FASTABS, que permiten al usuario realizar cruces de dos o más entradas y calcular una variedad de medidas estadísticas no paramétricas. CROSSTABS produce una secuencia de tablas de doble entrada que muestra en la dimensión vertical los valores de una variable y a lo largo de otra, los valores de la segunda variable. En el cuerpo de la tabla se muestran frecuencias condicionadas; éstas pueden expresarse como porcentajes del total del renglón, de la columna o de la tabla. El grado de asociación de dos variables, basado en la distribución de frecuencias, está dado por las siguientes medidas estadísticas: chi cuadrada, CRAMER, KENDALL, el estadístico gamma, SOMER y otros cuyo cálculo es optativo.

Otra técnica para el examen de relaciones entre dos o más variables en un formato de cruces la provee el sub-programa BREAKDOWN. Este procedimiento requiere que al menos la variable dependiente sea de escala ordinal. Calcula, además, promedios, desviaciones estándares y varianza para cada subgrupo considerado en una muestra. En este caso, cada promedio y desviación estándar resume la distribución de un renglón completo o de una columna de la tabla de cruces.

c) Análisis de correlación bivariada

En el análisis de correlación se provee al investigador de una técnica para la medida de relaciones lineales entre dos variables y produce una medida estadística, el coeficiente de correlación, que describe la fuerza de esa asociación.

El SPSS tiene dos sub-programas para calcular correlaciones: PEARSON CORR, que produce un coeficiente de correlación adecuado para datos con distribución normal de frecuencias; y NONPAR CORR, que permite al usuario calcular coeficientes de correlación de SPEARMAN o KENDALL, o ambos.

Las salidas de PEARSON CORR y NONPAR CORR son similares y proporcionan los coeficientes de correlación, el número de observaciones sobre el cual esa correlación está basada, y el nivel de significancia estadística de ese coeficiente. Además, cada procedimiento entrega en la salida matrices de correlación que se pueden utilizar como entrada en procedimientos posteriores.

d) Correlación y regresión multivariada

La correlación parcial y la regresión múltiple permiten al usuario efectuar una amplia variedad de análisis para explicar y predecir relaciones entre variables, cuando éstas satisfacen un mínimo de suposiciones de distribución y escalas requeridas para estas técnicas estadísticas.

La correlación parcial entrega un coeficiente que describe la relación lineal entre dos variables, mientras se ve afectada por otras. Se pueden considerar hasta cinco particiones para cualquier conjunto de variables y el usuario tiene un control total sobre las órdenes y las partes a ser consideradas. La salida de este sub-programa incluye coeficientes de correlación parcial, su nivel de significancia estadística y el número de casos sobre los cuales se basa cada etapa. También se pueden obtener correlaciones de orden cero, promedios y desviaciones estándares de las variables.

La regresión múltiple permite al investigador estudiar relaciones lineales entre un conjunto de variables independientes y una variable dependiente, mientras toma en consideración las relaciones entre las variables independientes. El objetivo básico de la regresión múltiple es producir una combinación lineal de variables independientes, la cual se correlacionará, tanto como sea posible, con la variable dependiente. Esta combinación lineal se puede usar para predecir valores de la variable dependiente y la importancia de cada una de las variables independientes en la predicción.

Se puede calcular la regresión sobre todas las variables a considerar, o usando una técnica "paso a paso", en la cual las variables son introducidas según un orden de prioridad indicado por el investigador.

La "salida" incluye los coeficientes de la ecuación de regresión, su error estándar, el nivel de significancia de los coeficientes, y otros valores que se calculan en cada etapa de la regresión.

## 4. Otras características importantes del SPSS

### a) Secuencia de los cálculos

El SPSS está dirigido, a través de sus distintan funciones, por una secuencia de tarjetas de control que el usuario debe preparar. Hay un programa de control en SPSS cuya única función es leer tarjetas de control, interpretarlas y hacer que se ejecute la función indicada en la instrucción.

El usuario debe ordenar las tarjetas de control en una secuencia apropiada, para lograr que el sistema tome las acciones que él desea; además, para que estas tarjetas sean reconocidas por el sistema, deben ser preparadas en un formato particular.

### b) Entrada y procesamiento de datos

Los datos se pueden introducir al SPSS en una variedad de formas; la más simple, y quizás la más común, es perforarlos en tarjetas e insertarlas junto con tarjetas de control que instruyen al sistema sobre el proceso.

Los datos están organizados, dentro del sistema SPSS, en unidades llamadas archivos SPSS ("SPSS System File"). Estos contienen los datos del usuario con su correspondiente "información". Cualquier archivo puede ser guardado permanentemente como archivo SPSS, en el medio de salida que el usuario prefiera. Estos archivos SPSS están en binario.

Las tarjetas de definición de datos entregan información sobre los datos a procesar. Una de estas tarjetas, la FILE NAME, da el nombre del conjunto de datos para futuras referencias. En la tarjeta VARIABLE LIST se indican los nombres de las variables que intervienen en el trabajo ("step"). Estos nombres de variables estarán permanentemente asociados con las correspondientes variables en el archivo, y todos los procesos futuros se efectuarán con referencia a esos nombres.

El tipo de variable y su ubicación en el registro de datos se especifica en la tarjeta INPUT FORMAT, y el número de casos a considerar, en la tarjeta # OF CASES.

El medio de entrada con que se han ingresado los datos se especifica en la tarjeta INPUT MEDIUM, mediance el uso de ciertas palabras clave.

La tarjeta PRINT FORMAT especifica el formato de impresión de las variables y se requiere sólo cuando hay variables en el archivo con caracteres no numéricos.

Estas seis tarjetas de control se requieren siempre cuando se define un archivo SPSS.

Los siguientes tres tipos de tarjetas entregan información adicional y son optativas.

La tarjeta MISSING VALUES permite al usuario designar hasta tres valores para cada variable, los cuales deben ser omitidos. Estos valores reciben un tratamiento especial en el análisis y cada sub-programa posee varias opciones que el usuario puede seleccionar para procesar esos valores "a omitir".

La tarjeta VAR LABELS permite asociar el nombre de una variable a otro más explicativo. Estos se imprimirán automáticamente en todas las tablas e informes donde intervengan las variables.

La tarjeta VELUE LABELS tiene una función idéntica, pero se refiere a valores individuales de las variables, o a clases.

La información contenida en las tarjetas de definición de datos puede "salvarse" permanentemente junto con los datos en un archivo SPSS.

Estos archivos SPSS se "salvan" mediante la tarjeta de control SAVE FILE. No se requiere una corrida especial para generar un archivo.

En general, un usuario puede ingresar sus datos desde cualquier tipo de archivo BCD; éstos corresponden a los que se han registrado mediante el esquema de registro BCD ("Binary Coded Decimal"), que es el usado normalmente para perforar tarjetas.

El sistema SPSS es instruído, en lo que se refiere a ejecución de cálculos estadísticos, por medio de un conjunto de tarjetas de "definición de tareas".

Las restantes cuatro tarjetas de control SPSS sirven simplemente para funciones específicas del sistema.

La tarjeta RUN NAME, que contiene cualquier texto, a gusto del usuario, identifica la corrida, y el contenido de ésta se imprime al principio de cada página del listado generado por la corrida.

La tarjeta READ INPUT DATA informa al sistema que el usuario ha terminado de definir el archivo: entonces interpreta la primera tarea estadística, y luego empieza a leer los datos dentro del computador.

La tarjeta FINISH simplemente informa al SPSS que la corrida ha terminado.

La tarjeta GET FILE hace que todos los datos y la información del archivo nombrado en esta tarjeta sean leídos por el computador, este archivo había sido creado previamente a través de un SAVE FILE.

c) Recodificación de datos

A fin de organizar los datos para el análisis, el usuario determina primero con cuáles variables desea tratar. El término variable se refiere a un cierto atributo que puede ser determinado o medido y debe ser diferenciado del término VALUE, que es el valor determinado o medido para una variable en particular.

Cuando los datos deben ser procesados por el computador, la manera como se proyecta esta codificación puede influir grandemente en la facilidad con que el computador lleve a cabo los cálculos que se desea. Frecuentemente, el sistema de codificación que el usuario ha utilizado para registrar sus datos no es el más conveniente para aplicar en todas las partes del análisis.

El usuario puede cambiar sus códigos después de entrados en su forma original al sistema. Los valores de todas las variables o de cualesquiera de ellas pueden ser cambiados por el usuario, mediante el proceso RECODE.

d) Transformación de variables

Las transformaciones pueden ser de dos tipos: incondicionales o condicionales. Las transformaciones incondicionales van definidas por la tarjeta de control COMPUTE, lo que origina una nueva variable en función de otras.

Las transformaciones condicionales se definen por la tarjeta IF: y permiten al usuario averiguar si cierta condición es verdadera y, en ese caso, la transformación se realiza. Podría usarse, además, una combinación de ellas para definir una nueva variable.

e) Muestreo, selección y asignación de importancia de datos (peso)

Se puede obtener una muestra aleatoría de los casos de un archivo, así como seleccionar casos específicos para procesarlos, o asignar un peso a ciertos casos. El usuario podrá

especificar todas las condiciones que desee para obtener la muestra, la selección o la asignación de peso a casos, durante cualquier corrida.

El usuario puede elegir el uso de tarjetas de control OPTIONS que le permitan producir los informes que le interesen en cada sub-programa.

Existen otras tarjetas que cumplen distintas funciones, relacionadas también con el análisis de los datos, pero su exposición haría aún más largo este resumen de las características del SPSS.

## 5. Condiciones operativas del SPSS.

Originalmente los programas SPSS fueron elaborados para ser usados en una máquina con las siguientes características: Una unidad central de 262.144 Bytes de almacenamiento principal; unidades de disco tipo 2314, y cintas de 9 o 7 canales de grabación. La mayoría de los programas están escritos en FØRTRAN IV, teniendo algunas rutinas de entrada/salida en lenguaje ensamblador 360. A esta versión se conoce como SPSSH. Se han efectuado las adaptaciones correspondientes, a fin de hacer funcionar este sistema en computadores de la serie CDC 6000.

El sistema disponible en CELADE es una versión restringida del anterior, conocido como SPSSG, y es operable en un equipo de las siguientes características:

- Unidad central de proceso IBM 360/40-G con 128 KBytes de memoria principal.
- 3 Unidades de disco tipo 2311
- 4 Unidades de cinta tipo 2 400, de 9 canales con 800 BPI
- Lectora y perforadora de tarjetas (2501 y 1442)
- Impresora tipo 1403-N1

y bajo la versión 19.6 del sistema operativo (O.S.) de IBM.

A muy corto plazo se espera tener funcionando la versión SPSSH en un computador con las características mencionadas.

# IV. DESCRIPCION DE LOS PROGRAMAS OSIRIS/40

## ("Organized Set of Integrated Routines for Investigation with Statistics")

### 1. Introducción

OSIRIS es un juego de programas elaborados en el "Institute for Social Research (ISR)" de la Universidad de Michigan, para análisis de datos en ciencias sociales, especialmente datos de encuestas. Comprende un paquete de programas muy comprensible y de fácil uso.

Existe una amplia gama de programas disponibles. La posibilidad de manejo de informaciones incluye la corrección de archivos de datos y la transformación lógica y aritmética de ellos. La capacidad de análisis de datos estadísticos comprende una variedad de programas de análisis multivariados y no paramétricos.

Los programas, que han demostrado ser de fácil uso, forman un paquete integrado; por ejemplo, la salida de los programas usados en una etapa del análisis, es compatible con los requisitos de entrada de los programas usados en etapas posteriores. Además, las convenciones para el uso de programas son estándares. A veces resulta deseable excluir ciertos casos del análisis. La mayoría de los programas OSIRIS/40 permite esa opción, si se desea, para cada programa; la subdivisión está especificada exactamente en la misma forma. De igual modo, todos los programas de análisis cuentan con un mecanismo de recodificación temporario, mediante el cual se pueden combinar variables y construir nuevas variables para un análisis específico.

Los archivos de datos no necesitan preparación, antes de utilizar un programa OSIRIS. Tampoco hay mayores restricciones en cuanto a la forma de almacenar los datos, a la extensión de los registros o el número de registros por caso. Estos pueden estar en tarjetas, en cintas o en discos. El único requisito es que los datos deben almacenarse secuencialmente, en forma de caracteres, en registros de largo fijo, con el mismo número de registros para cada caso. Los programas de

análisis de OSIRIS también pueden ser usados con archivos de datos en binario (punto fijo o flotante).

Un importante mecanismo de OSIRIS es el uso de un archivo de descripción de los datos, ("diccionario"), de fácil preparación. Los registros del archivo de descripción de datos contienen la información sobre cada una de las variables en estudio. El archivo de "diccionario", junto con el archivo de datos que éste describe, conforma un juego de datos auto descritos conocido como un "juego de datos OSIRIS". Una vez preparado el diccionario, éste puede mantenerse en tarjetas, cintas o discos, y ser usado cuantas veces se desee.

Si se desea, el diccionario puede ser confeccionado cada vez que se ejecuta un programa OSIRIS, incluyendo solamente, los registros de las variables que realmente serán usados en esa ocasión.

Los "juegos de datos OSIRIS" pueden ser preparados en forma relativamente permanente, con un programa de confección de archivo que, además, realiza ciertas funciones de verificación y confrontación. No obstante, dicho proceso no es condición previa para usar los programas de análisis de OSIRIS, aunque es recomendable para aquellos archivos de datos que serán consultados frecuentemente.

El uso de los "juegos de datos OSIRIS" sólo requiere un número de variables para identificarlo. Por ejemplo: se dispone de un nombre variable para "rotular" la salida impresa, para facilitar la interpretación; la ubicación de la variable dentro del registro de datos es recobrada para hacerla accesible durante el proceso. Los valores del código designados "datos faltantes" son identificados para asegurar su tratamiento adecuado durante el análisis.

En las páginas siguientes se presenta una breve descripción de algunos programas de OSIRIS/40. 3/

3/ Para mayores informaciones véase el
"Manual de Usuarios de OSIRIS/40
Institute for Social Research (ISR)
University of Michigan
P.O. Box 1248, Ann Arbor
Michigan 48106, U.S.A.

## 2. Programas de Análisis

### a) Cuadros univariables y bivariables (TABLES) PG625

El programa produce cuadros de frecuencias ponderadas y no ponderadas, con porcentajes opcionales de fila, columna y esquina. Para los cuadros univariables computa la media, moda (la primera si hubiera más de una), la mediana, varianza (imparcial), desviación estándar, "skewness" y "kurtosis". Para los cuadros bivariables las medidas opcionales son pruebas-T de medias, (medidas independientes no repetidas) entre cada par de filas; chi-cuadros; coeficiente de contingencia C; "Cramer's V; Tau-A, Tau-B, Tau-C: gamma, lambda-a (columna por fila), lambda-b (fila por columna) y lambda.

### b) Análisis de varíanza de una dirección (F-MEANS) PG640/PG650

El análisis de varianza de una dirección analiza hasta 99 grupos. Los grupos se definen por valores de una variable de control. El programa calcula medias y desviaciones estándares para cada celda, sobre la variable dependiente, y realiza un análisis de varianza a través de las celdas.

### c) Correlación de datos faltantes (PDC) PG660

El programa de correlación de datos faltantes produce una matriz de correlaciones producto-momento de Pearson entre todos los pares de variables, en una lista proporcionada por el usuario. El programa permite datos faltantes al calcular separadamente el coeficiente para cada par de variables, basándose en el sub-conjunto de casos con resultados válidos en ambas variables. Siempre que se imprime una matriz de correlación también es posible obtener como salida en tarjeta o cinta magnética la matriz de correlación que podría servir de entrada para otros programas de OSIRIS.

### d) Correlación con sub-conjunto (STRIPCØR) PG665

Similar al programa de correlación de datos faltantes (anterior), STRIPCØR tiene, además, una opción para sacar una matriz rectangular, de manera que si se desean las correlaciones de algunas variables con muchas otras, conviene correlacionar con sub-conjunto.

e) Correlación Punto Biserial (PBSCOR) PC690

Este programa calcula e imprime coeficientes de correlación biseriales o puntos biseriales entre variables dicotómicas y continuas.

f) Estadísticas no paramétricas (NPSTAT) PH710

El programa obtiene e imprime estadísticas basadas en rangos, de acuerdo a la solicitud del usuario: prueba "Mann-Whitney"; prueba de signo; prueba "Wilcoxon" y/o coeficientes de correlación de rango "Spearman".

g) (THAID) PH715

Análisis multivariado de comportamiento de variables dependientes de escala nominal y/o variables de escala ordinal. El programa emplea una división binaria, con estadística "criterio THETA" (definida como la proporción de la muestra en la moda).

h) Regresión Lineal (REGRESSN) PH720

El programa realiza regresiones estándares lineales en etapas. La entrada puede constar de un "Modem" sin datos faltantes sobre las variables deseadas, o una matriz de correlación como la salida entregada por el programa de correlación de datos faltantes, o el programa de correlación con sub-conjunto.

i) Correlación Parcial (PARCOR) PH730

El programa de correlación parcial genera matrices de coeficientes de correlación parcial para variables seleccionadas, sobre un conjunto de datos. La entrada al programa es una matriz de correlación, como aquellas que generalmente son obtenidas por el programa de correlación de datos faltantes, el programa de regresión o el programa de correlación con subconjunto. El usuario puede especificar las sub-matrices de la matriz de entrada. Para cada sub-matriz, además de los coeficientes de correlación parcial, el programa imprime: i) la matriz inversa; ii) la correlación múltiple de cada variable en el sub-conjunto utilizando todas las variables restantes; y iii) los coeficientes estandarizados de regresión.

j) Agrupación Jerárquica (HICLSTR) PH735

Es un programa de agrupación jerárquica basado en un algoritmo propuesto por Johnson (Psychometrik 1967,32,241-254). La entrada al programa consta de una matriz de proximidades. La matriz de coeficientes de correlación, generada por otros programas de OSIRIS, puede servir de entrada.

k) Análisis multivariable de varianza (MANOVA) PH740

Es un programa de análisis multivariable de varianza para analizar hasta 8 factores. Los números de celdas no necesitan ser iguales. El análisis de co-varianza es posible hasta con 8 co-variantes. (MANOVA originalmente fue desarrollado en el Laboratorio Biométrico de la Universidad de Miami).

l) Prueba T Tri-Variable (TRIVAR) PH750

El programa realiza pruebas "T de STUDENT" entre medias de grupos de una variable dependiente. Una o dos variables de control pueden ser usadas para definir los grupos. Si se utilizan dos variables de control se computan 6 pruebas T; una estadística T para la diferencia de las medias de los dos grupos dentro de cada predicción para cada grupo de la otra variable pronosticada.

m) Detección automática de interacción (AID) PH760

AID es algo parecido al programa de "regresión paulatina", en que las variables independientes "pronosticadas" no necesitan ser cuantitativas. Las suposiciones lineales y los aditivos inherentes a las técnicas convencionales de regresión múltiples no son requeridas.

El objeto del programa es explicar la varianza de la variable dependiente. El programa produce una serie de varianzas mediante la división sucesiva en sub-conjuntos. Cada selección de variable independiente y su punto dividido, se basa en el criterio de que proporciona la reducción máxima posible de varianza en la variable dependiente.

n) Detector automático de interacción III (AIDIII) PH763

AID III es parecido al AID, excepto que es posible maximizar las diferencias, no sólo en medias de grupos, sino también

en declives o líneas de regresión. Además, es posible hacer que el programa mire hacia adelante, es decir, examine el poder explicativo de una, dos o tres divisiones sucesivas, antes de seleccionar "la mejor".

a) Análisis de clasificación múltiple (MCA) PH770

Es un programa para regresión múltiple, utilizando pronósticos categóricos y en el que la variable independiente puede constar de escalas nominales. La salida comprende un coeficiente de correlación múltiple y la suma de cuadrados.

## 3. Programas de uso general

a) Impresión de "Rótulos" estándares (LABPRT) PA002

Este programa lee rótulos ("label") en cinta magnética e imprime el contenido de cualquiera o de todos los rótulos estándares, si los hubiera. Su objetivo principal es detectar las discrepancias en el procesamiento de rótulos.

b) Listador/Reproductor de tarjetas (PUNLIST) PC200

El programa hace listados y reproduce tarjetas o archivos en imagen de tarjetas.

c) Copia múltiple (NCØPY) PC205

El programa hace copias múltiples de lotes de tarjetas.

d) Programa de listado de conjunto de datos OSIRIS (DSLIST) PC210

Programa de listado para imprimir juegos de datos OSIRIS. El usuario puede especificar un subconjunto de variables y/o casos para impresión.

e) Traspaso de cinta a tarjeta o cinta (TCØT) PC220

El programa acepta como entrada un juego de datos OSIRIS y saca tarjetas o un archivo de imagen de tarjeta. Generalmente es utilizado para hacer formatos de datos, de manera

que éstos puedan ser utilizados en programas distintos a aquellos que conforman los programas de OSIRIS, los que requieren como entrada registros en imagen de tarjeta.

## 4. Programas de manejo de datos

### a) Verificación, comparación (MERCHECK) PAC15

El programa intercala y verifica datos clasificados para encontrar tarjetas faltantes y/o extrañas. Además, elimina tarjetas inválidas y/o casos cuyas tarjetas faltan. Las entradas son archivos en forma de tarjetas o imagen de tarjeta en cinta o discos. La salida de un archivo depurado en cinta o disco, y un listado ("print-out") que señala todos los errores de intercalación en los datos. Este programa generalmente es ejecutado antes del programa de creación de archivo (PA020)

### b) Creación de archivo (BUILD) PA020

El programa convierte los datos almacenados en tarjeta, o cinta en imagen de tarjeta, en un juego de datos estándar OSIRIS que consta de un diccionario y un archivo de datos. Los datos son verificados para encontrar caracteres no numéricos y en blanco, y puede realizarse una verificación ("edit") limitada. La salida del programa de creación de archivo puede servir de entrada para los programas OSIRIS de análisis. La creación del archivo estándar se hace solamente una vez, para un conjunto de datos.

### c) Confrontación de códigos inválidos (WCC) PD510

El programa de confrontación de códigos inválidos explora las variables en un conjunto de datos OSIRIS, para hallar valores no admisibles. El usuario puede especificar los códigos válidos y/o inválidos, para cada variable que él desee verificar. Alternativamente, si los códigos de las variables fueron especificados al formarse el archivo, los códigos consignados en el diccionario de códigos pueden ser utilizados para cotejar, sin mayores especificaciones.

### d) Corrección de archivo (TCØR) PE410

Este programa, usado para corregir datos OSIRIS, permite: corrección de variables específicas dentro de un registro

de datos; eliminación de registros específicos: listado de registros específicos: corrección de registros de diccionario. Es usado cuando se descubren errores en los datos, después de haberse creado el archivo de datos OSIRIS (estándar).

e) Subdivisión de matriz (SUBMAT) PE425

El programa de subdivisión de matriz crea subconjuntos de matrices existentes para ser usados en otros programas de análisis. Es usado como un paso intermedio entre programas como el de correlación de datos faltantes, que produce una matriz de correlación de hasta 200 x 200 variables y programas de análisis, tales como el de correlación parcial y el de regresión lineal, que no aceptan matrices tan grandes.

f) Copia de archivos y subconjuntos (FILECOPY) PE435

Es un programa general, para copiar casos seleccionados y variables, en un nuevo archivo estándar OSIRIS.

g) Intercalación de Archivos (MMP) PE 440

Es un programa para intercalar y/o combinar dos juegos de datos OSIRIS. La salida es un nuevo juego de datos estándares OSIRIS. El usuario especifica las variables de identificación que se comparan en los dos archivos de entrada: hay varias opciones para los registros no comparados, incluyendo el relleno de registros con casos faltantes o la eliminación del caso. También se especifican las variables de cada archivo de entrada que se va a transferir al archivo de salida. Así, el usuario puede controlar tanto los casos como las variables que conforman el nuevo juego de datos. Es un programa bastante flexible, ya que la intercalación tiene muchas aplicaciones. Puede servir para agregar nuevos datos a un archivo existente, o para agregar datos constantes en un grupo de tarjetas perforadas.

h) Construcción y registro de Indices (ICON) PF510

ICON es un programa de manejo de datos y construcción de índice que permite prácticamente cualquier tipo de registro y transformación de variable. La salida consiste en un nuevo juego de datos OSIRIS, generados por operaciones indicadas por el usuario. Las operaciones disponibles pertenecen a cuatro categorías generales: 1) transferencia sin alteración al

nuevo juego de datos; 2) generación de variables; 3) cómputos aritméticos: y 4) escritura de un registro de salida. El usuario puede especificar instrucciones "lógicas" para el flujo directo del programa. ICÓN es el programa más importante de manejo de datos de OSIRIS. Se usa para crear las variables reales de datos no depurados que el investigador desea utilizar en análisis posterior.

i) Agregación (ACCREG) PF520

Este programa adiciona datos recopilados a un nivel o unidad de análisis, a otro nivel diferente. Las variables se pueden agregar usando una variedad de estadísticas. La salida del programa de agregación es un nuevo juego de datos OSIRIS, en el cual cada registro contiene los datos resumidos para cada sub-conjunto y/o nuevo juego de datos OSIRIS, con un registro por caso entrado, conteniendo variables originales junto con índices de resumen.

El programa es útil para aplicaciones tales como el resumen de transacciones dentro de los departamentos o medios de cómputo, y cuentas de datos que describen números de niños en una clase. Se dispone de capacidades muy flexibles de subconjunto para cada estadística y cada formato de salida.

j) Cuartiles (NTILE) PF530

El programa NTILE genera la función de distribución, "función Lorenz", y la prueba "Kolmogorov-Smirnov", para variables especificadas por el usuario. Los puntos de división en sub-intervalos son impresos y opcionalmente también las funciones "Lorenz" son calculadas empleando interpolación lineal. En vista de que el usuario puede especificar el número de intervalos, el programa puede ser usado para divisiones de medias, cuartiles, deciles, etc.

k) Resumen (SUMMAR) PF540

Este programa realiza funciones similares a las de agregación, excepto que computa las mismas estadísticas sobre una lista de variables, en vez de tratar separadamente a cada una. La salida consiste en un nuevo juego de datos OSIRIS que contiene los datos resumidos correspondientes a cada sub-conjunto, además de la salida impresa. El programa es fácil de usar, ya que los formatos e ítems del diccionario tienen su juego de omisión.

## 5. Servicios Generales de OSIRIS

### a) Programa de Control de OSIRIS (ISRSYS)

La mayoría de los programas, incluyendo todos los de análisis en OSIRIS, pueden ser cargados y ejecutados mediante un programa especial de control ISRSYS. Las tarjetas de control suministradas normalmente por el usuario, y varias opciones, como capacidades, por ejemplo, son proporcionadas automáticamente.

### b) Dispositivo OSIRIS de recodificación (RECODE)

Mediante el uso del RECODE, y basándose en las variables ya existentes, fácilmente pueden ser construídas nuevas variables para uso en la ejecución específica de un programa de análisis.

### c) Dispositivo de filtro OSIRIS

Si no es necesario utilizar los archivos de datos en su totalidad, es posible usar un sub-conjunto de casos, en un programa de análisis OSIRIS. Todos los programas tienen una capacidad de filtro, mediante la cual, con una tarjeta estándar de control, se elige los casos necesarios para una corrida específica, seleccionando automáticamente estos casos de un archivo maestro de datos.

### d) OSIRIS/40 - El ambiente operativo

Originalmente, los programas OSIRIS'40 fueron elaborados para ser utilizados en una máquina de la siguiente configuración: Una unidad central de 131.072 bytes de almacenamiento principal y el juego universal de instrucciones (aritmética de punto fijo y aritmética de punto flotante): cuatro unidades de disco 2311; cuatro unidades de cinta de 9 canales; una impresora; una lectora de tarjetas y una perforadora. La configuración mínima consistiría en una máquina con por lo menos 106 K. bytes fuera del núcleo, dos mil pistas de 2311 de espacio en disco, y suficiente capacidad de disco y cinta para actualizar los archivos.

Actualmente, los programas en CELADE se están ejecutando en un sistema IBM-360/40 con 128 K Bytes y 3 discos 2311;

4 cintas, lectora y perforadora e impresora; bajo la versión N° 19.6 del Sistema Operativo (O.S.) de IBM como programa de control primario. Aunque la mayoría de los programas se efectúan en una partición de 100 K., se ha establecido que los programas grandes (MCA, HANOVA), requieren una partición de 130 K.

OSIRIS/40 fue diseñado para aprovechar el Sistema Operativo proporcionado por IBM. Las cintas y los discos son usados intensivamente por OSIRIS, ya que éstos dispositivos de almacenamiento son bastante eficientes.

Los programas IBM de uso general, particularmente el programa de clasificación/intercalación, aportan capacidad adicional al OSIRIS/40. El ISR OSIRIS/40 es suplementado con dos programas diseñados y suministrados por la "Cambridge Computer Association, Inc.," Cambridge, Massachussets. Estos son "CROSSTABS II", un programa de tabulación muy flexible, y "UTILITY - CODER/360", un lenguaje de procesamiento y programación para manipulación de datos.

# V. DESCRIPCION DE LOS PROGRAMAS MINI-TAB

## 1. Introducción

Los programas MINI-TAB están diseñados para usarse en computadores con memoria pequeña, con el propósito de tabular datos de encuestas cuando no estén disponibles computadores de gran porte o cuando no se cuenta con programas más elaborados. El MINI-TAB ha sacrificado embellecimiento de la presentación de resultados para ganar versatilidad. Los programas MINI-TAB escritos en "FORTRAN" básico, ejecutan algunas operaciones más lentamente que los programas escritos en lenguajes de niveles más bajos. Sin embargo, el programa presenta la gran ventaja de poder ser utilizado en casi todos los computadores dignos de ese nombre, sin importar el tamaño, edad o fabricante de la máquina. Hay limitaciones, por supuesto, pero donde haya un compilador FORTRAN y unas pocas posiciones de memoria disponibles, los programas MINI-TAB pueden usualmente adaptarse con un mínimo de dificultad. Aunque los programas hayan sido escritos originalmente en FORTRAN IV, las limitaciones del FORTRAN II fueron consideradas para facilitar la conversión. Además, los programas no presuponen la disponibilidad de equipos tales como unidades de cinta y disco, aunque esté previsto el uso de estas unidades cuando estén disponibles.

El tamaño del computador no limita el número de casos (el tamaño de la muestra) que puede procesarse. Los programas MINI TAB pueden procesar estudios con varios miles de casos. La memoria del computador limita el número de estadísticas(variables, campos de datos y tablas) que pueden procesarse en una corrida. Sin embargo, estas limitaciones pueden superarse mediante la división de las tabulaciones en secciones, y procesando cada sección en corridas separadas. Por ejemplo, se puede procesar cincuenta tablas en una corrida del "MINI TAB TABLES y cuarenta tablas más en una corrida posterior. Las limitaciones exactas en el número de estadísticas permitidas dependerán de la manera en que sean requeridas las tabulaciones y del tamaño del computador.

La simplicidad de los programas MINI TAB puede ser una ventaja, tanto para el programador avanzado como para el principiante. Aún un individuo poco familiarizado con el procesamiento de datos puede entender el sistema MINI TAB, con unas pocas horas de estudio y práctica. MINI TAB también introduce a varios aspectos de la programación en FORTRAN, que

serán potencialmente de utilidad en el futuro. Para los programadores avanzados, para quienes el lenguaje FORTRAN ya es conocido, MINI TAB es un instrumento que podrán dominar y ajustar a sus propias necesidades. A diferencia de los programas que consisten de varios cientos de instrucciones, la simplicidad y tamaño de los programas MINI TAB invitan a adiciones y opciones especiales diseñadas por el programador.

## 2. Tipos de información que el computador requiere del usuario

Hay tres tipos de información que el usuario debe proveer al computador, en cada trabajo que realice en MINI TAB. Estos están descritos a continuación, en el siguiente orden: a) Tarjetas del Sistema; b) Programa fuente; y c) Datos.

### a) Tarjetas del Sistema

Son tarjetas específicas de cada instalación particular del computador en que el usuario esté trabajando. Las tarjetas del sistema son pocas, en número, y a menos que se utilicen opciones para la entrada o salida por medio de cinta o disco, son sencillas. Las tarjetas del sistema consisten primeramente de una tarjeta de "JOB", que alerta al computador del comienzo de un nuevo programa; segundo, una tarjeta que llama al compilador FORTRAN necesario; tercero, las tarjetas que definen el comienzo y el fin del programa, y cuarto, las tarjetas que definen el comienzo y el fin de los datos. Las tarjetas del Sistema son necesarias para la corrida de cualquier programa en FORTRAN, en la instalación del computador, y si el usuario no está familiarizado con estas tarjetas de su instalación particular, puede, generalmente, recibir asistencia del personal local en la preparación de las mismas.

### b) Programa Fuente

El segundo tipo de información necesaria para el proceso MINI TAB es un programa fuente MINI TAB. El programa se compone de una serie de instrucciones codificadas en lenguaje de computador FORTRAN. Por medio de un compilador FORTRAN, el computador debe traducir las instrucciones del programa a un nivel de instrucciones más bajo, aplicable sólo a determinados computadores. Las instrucciones de fuente del SISTEMA MINI TAB están agrupadas de acuerdo con sus funciones en programas (Módulos) que están interrelacionados, pero pueden usarse

independientemente uno de otro. El uso de los módulos independientemente uno de otro, ahorra memoria en muchas instalaciones pequeñas, puesto que el proceso de sobreposición (usar la memoria del computador sólo para la parte del programa a un tiempo) o es difícil de encontrar o es problemático. El Capítulo 3, describe el MINI TAB EDIT, un módulo independiente de limpieza y corrección de datos. El programa MINI TAB EDIT puede ser usado para localizar los códigos fuera de rango y las inconsistencias lógicas dentro de los datos, antes de analizarlos. El Capítulo 4, está dedicado a MINI TAB FREQUENCIES, un programa para el cálculo de frecuencias marginales y la descripción de las variables, una a una. El Capítulo 5 describe el MINI TAB TABLES, un programa para hacer tabulaciones cruzadas.

c) Datos

Los datos son leídos de acuerdo con las instrucciones del programa, y consisten de dos partes: Tarjetas de control, y Unidades de estudio.

Los datos de las unidades de estudio (o casos, como también se denominan) están organizados de acuerdo con la siguiente jerarquía:

Estudio
  Grupo mayor
    Unidades de estudio (casos)
      Campos de datos

d) Adquisición de los Programas MINI TAB

Los programas fuentes de MINI TAB son gratuitos. Puesto que los programas están listados en el Manual "MINI TAB EDIT, MINI TAB FREQUENCIES y MINI TAB TABLES: JUEGO DE TRES PROGRAMAS ESTADISTICOS INTERRELACIONADOS PARA COMPUTADORES PEQUEÑOS" Henry G. Elkins, Universidad de Chicago, Community and Family Study Center, N° 7, cualquiera puede simplemente perforar las instrucciones usadas en los programas. El "Community and Family Study Center", por un pago nominal de la hoja de procesamiento, manipulación y costos de correo, enviará copias de los programas fuentes, en forma de tarjetas perforadas, a cualquier lugar del mundo que lo requiera.[4/]

4/ Para obtención de copia del programa, escribir a: Director of Computer Applications, Community and Family Study Center, University of Chicago, Chicago, Illinois, 60637, U.S.A.

## 3. "MINI TAB EDIT", un programa de limpieza de datos

### a) Introducción

La limpieza de datos es un pre-requisito esencial para etapas posteriores de análisis, si se desean resultados seguros. Frecuentemente se observa un gran número de errores que pueden ocurrir en la codificación, perforación o manipulación de tarjetas. Muchos de estos errores pueden corregirse mediante pruebas en las que se observan los valores que están por fuera del rango permitido y mediante pruebas de consistencias lógicas en los datos o a través de relaciones que se deben obtener entre valores de códigos. En vez de programar cada control de error, un proceso costoso en tiempo y dinero, MINI TAB provee una serie de controles de consistencia y de rangos que trata las necesidades más generales de limpieza.

El programa MINI TAB, en virtud de ser adaptable a computadores pequeños, permite la detección de códigos errados y consistencias lógicas cerca de la fuente de error. Donde se está realizando el estudio, los formularios originales de entrevista y las hojas de codificación pueden contribuir a correcciones más precisas. Por otra parte, la información concerniente a los tipos de errores cometidos durante la codificación y perforación pueden ayudar a prevenir errores en el futuro.

En caso de que esté presionado por el tiempo, el investigador puede desear suprimir el proceso de limpieza. Uno de los problemas en suprimir el proceso de limpieza es que además del error de muestreo y algún tipo de respuesta sesgada, no se puede estimar fácilmente la cantidad de errores resultantes de la codificación o perforación; y antes que se logre analizar a mano una muestra adecuada, para determinar la extensión de los errores de codificación y perforación, puede tenerse controlada la totalidad de los casos, por computador. Siempre se puede cambiar los códigos fuera de rango a un código de "no respuesta"; pero puede considerarse esto como un proceso de supresión, en vez de corrección. MINI TAB no propone correcciones automáticas de errores. En su lugar, el programa simplemente lista los errores, de modo que el investigador pueda tomar sus propias determinaciones, a la luz de los documentos originales. Si los documentos originales no están disponibles, o si el investigador no quiere emplear tiempo haciendo las correcciones, el programa MINI TAB EDIT por lo

menos provee un instrumento para juzgar la calidad, tanto de los datos como de las conclusiones.

b) Tipos de errores de control

El programa provee seis tipos de control de errores. El primero es un control para los rangos de valores; el segundo, tercero y cuarto, son llamados controles aritméticos; el quinto y sexto son controles lógicos.

Control de rango

(1) $Min \leq A \leq Max$ — El programa prueba si el valor en un campo dado A está dentro de un rango o es igual a un mínimo y a un máximo especificado por el usuario. Se permiten hasta cinco excepciones para cada campo de datos, esto es, para valores que están fuera de rango, pero que no obstante son válidos.

Controles aritméticos

(2) $A > B$ — El programa verifica si el valor encontrado en el campo de datos A es mayor que el valor en el campo de datos B.

(3) $A \geq B$ — El programa controla si el valor en el campo de datos A es mayor o igual al valor en el campo de datos B.

(4) $A = B$ — El programa prueba si el valor en el campo de datos A es igual al valor en el campo de datos B.;

Para los errores mencionados en (2), (3) o (4), el usuario puede especificar por cada campo de datos hasta cinco valores de excepción. Por ejemplo, un investigador puede establecer que la edad al casarse debe ser superior a cierto número de años, exceptuando los códigos de no respuesta, no sabe, o "no se aplica" (soltero). Es posible diseñar controles aritméticos más complejos, usando la SUBROUTINE INDEX, y crear nuevos campos de datos para la unidad de estudio leída.

Controles Lógicos

| | |
|---|---|
| (5) Si A = X entonces B = Y | El programa verifica que si el campo de datos A es igual al valor X, entonces el campo de datos B es igual al valor Y. |
| (6) A = X si y solo si B = Y | El programa verifica que si el campo de datos A es igual al valor X, entonces el campo de datos B es igual a Y, entonces el campo de datos A debe ser igual a X. |

Para los controles lógicos el usuario especifica los campos primarios y secundarios (A y B) más los valores X e Y para cada campo. Considerando que en muchas situaciones el usuario puede querer especificar más de un valor para cada X o Y, el programa permite hasta cinco valores a ser especificados para X é Y. Por ejemplo, el error tipo cinco puede tomar la siguiente forma: Si el campo de datos A es igual a $X_1$, o $X_2$, o $X_3$, o $X_4$, o $X_5$ entonces el campo de datos B debe ser igual a $Y_1$, o $Y_2$, o $Y_3$, o $Y_4$, o $Y_5$.

## 4. "MINI TAB FREQUENCIES". Programa de descripción de datos con una entrada

Después de la corrección o limpieza de los datos, y antes de efectuar tabulaciones cruzadas o análisis, la mayoría de los investigadores prefieren examinar los datos en la forma de marginales. MINI TAB FREQUENCIES provee un medio rápido de cálculo de frecuencias, porcentajes y porcentajes acumulados.

Una opción importante a menudo especificada es la exclusión de categorías no aplicables para los cálculos estadísticos. Un problema, en muchos programas de frecuencias para computador, es que las categorías no aplicables, tales como códigos de "no respuesta", distorsionan los porcentajes, promedio y desviación estándar y otros indicadores estadísticos calculados. Por ejemplo, el promedio del campo de datos "Edad del Matrimonio" tiene poco sentido si la frecuencia para la categoría "soltero" es incluída en los cálculos. A menos que el usuario no especifique lo contrario, MINI TAB FREQUENCIES calculará las estadísticas basado en todos los valores encontrados pero el programa provee la opción de excluir de los cálculos

estadísticos hasta siete valores no aplicables para cada campo de datos. Las frecuencias de estas categorías no aplicables son impresas, pero ellas no son tomadas en cuenta para los cálculos de porcentajes y otros estadísticos.

## 5. "MINI TAB TABLES". Un programa para Tabulaciones cruzadas

El programa MINI TAB TABLES provee un método flexible de tabulaciones cruzadas que es a la vez simple de entender y fácil de usar. Como un procedimiento normal, el programa provee la recodificación simple y el reagrupamiento de valores para la construcción de variables con los campos de datos originales. La SUBROUTINE INDEX puede ser usada para recodificaciones más complejas. Para cada variable se pueden definir hasta cien categorías, dando el rango de los valores y los nombres de las categorías. El usuario asigna a cada variable un número único de identificación, el que puede citar posteriormente, para combinar hasta nueve variables en una tabla. El número de identificación de la variable no necesita cambiarse de una corrida del programa a otra; por consiguiente, el usuario puede asignar a sus variables números permanentes, que son independientes de la secuencia en que se definen las variables y del número de variables definidas en cada corrida. Las frecuencias de las casillas y los totales en las tablas pueden ser tanto ponderadas como no ponderadas. Como parte del procedimiento, se calculan los porcentajes sobre las columnas para las casillas y los porcentajes marginales y totales. Además, el programa calcula la estadística del Chi-cuadrado y varias medidas estadísticas de asociación, incluyendo el Gamma de Goodman y Kruskal y, para tablas de tres o más variables, el Gamma parcial de Davis

## 6. Otras posibilidades de MINI TAB

### a) Uso de cintas magnéticas

Las cintas magnéticas tienen la ventaja de facilitar el uso de grandes cantidades de datos, y a una velocidad mayor que la que se logra con las tarjetas perforadas. En una aplicación típica, el usuario puede querer leer los datos de tarjetas en el programa MINI TAB FREQUENCIES y utilizar la opción de grabar los datos en una cinta que será utilizada

posteriormente en las corridas del programa MINI TAB TABLES o en otros programas. Se considera que todas las cintas grabadas o leídas en el sistema MINI TAB estarán escritas bajo un FORMAT (cintas en binario o sin FORMAT no son aceptables en el sistema MINI TAB). Así, las cintas grabadas por otros programas diferentes de MINI TAB y que van a ser leídas por uno de los programas MINI TAB, no deben estar en binario. Además, el usuario tiene que proveer, en las tarjetas de control del Sistema, las especificaciones apropiadas del archivo: longitud del registro lógico, tamaño del bloque, etc.

Puesto que las cintas no están limitadas a una imágen de tarjeta arbitraria de ochenta columnas, es importante especificar la longitud del registro. Además, en computadores grandes, es posible agrupar ("block") los datos en la cinta. Las cintas agrupadas ("blocked") tienen la ventaja de comprimir más datos en menos pies de longitud, omitiendo los saltos entre los registros para cada grupo de estos. El tamaño del grupo es conocido como tamaño del bloque, y se especifica en las tarjetas de control del Sistema. Obviamente, mientras menos espacios en blanco haya entre los datos, menor será el tiempo en recorrer la cinta.

Como las tarjetas de control del Sistema para el uso de cintas son especiales para cada instalación de computador, el usuario que piense usar cintas debe familiarizarse con las instrucciones del sistema local, para el manejo de archivos de datos en FORTRAN.

b) Como construir índices y modificar campos de datos

Los programas MINI TAB proveen una opción de construcción de campos de datos especiales y transformaciones matemáticas, o combinaciones de datos de varios campos de datos originales, cuando estas transformaciones están fuera del proceso normal de recodificación del programa MINI TAB TABLES. Todas las modificaciones e índices deben construirse por medio de sentencias FORTRAN, pero sólo es necesario un conocimiento mínimo de FORTRAN, en la mayoría de las modificaciones e índices que se construyen. Todos los programas MINI TAB tienen incluída una SUBROUTINE llamada INDEX. Originalmente, esta SUBROUTINE es falsa. El arreglo K en la SUBROUTINE está dimensionado en común con el arreglo K del programa principal. Por lo tanto, cualquier alteración hecha en un miembro del arreglo

K en la SUBROUTINE es una alteración automática en los datos de entrada. La ventaja de construir los índices con una SUBROUTINE es que esta puede compilarse y corregirse separadamente del programa principal. Además, no hay posibilidad de duplicar los números de las instrucciones usadas en el programa principal.

Por ejemplo, si se lee 50 campos de datos de tarjetas, se puede querer construir un índice que resuma los campos de datos 11 a 15

K(51)=K(11)+K(12)+K(13)+K(14)+K(15)

En el ejemplo anterior K(51) está al lado izquierdo de la ecuación. Por lo tanto, la suma de los campos K(11) hasta K(15) está almacenada en la posición 51 del arreglo de datos K. La suma está ahora disponible en el programa principal, y puede usarse como una fuente de datos para la construcción de variables en el programa MINI TAB TABLES o para obtener los marginales en el programa MINI TAB FREQUENCIES.

c) Cómo procesar datos con un número variable de campos de datos por Unidad de Estudio

Algunas veces el investigador necesitará procesar datos cuyas unidades de estudio (casos) no tienen un número fijo de campos de datos. Por ejemplo, un estudio de mujeres puede incluir dos tarjetas para todas las mujeres sin importar el estado marital, y una carjeta adicional solo para las mujeres actualmente casadas. Puesto que los programas MINI TAB presumen que cada unidad de estudio tiene un número fijo de campos de datos, se aconsejan a continuación algunos métodos para superar esta restricción.

Opción número uno

Preparar tarjetas falsas, o campos de datos falsos (o en cinta, imágenes de tarjetas falsas), para tener todos los casos con el máximo número de tarjetas. Cuando se tiene un número pequeño de casos esto puede hacerse fácilmente, incluyendo tarjetas blancas. Sin embargo, cuando el número de unidades de estudio es grande, ese procedimiento puede representar un aumento substancial de trabajo.

En ese caso, puede buscar la inclusión de estas tarjetas por medio de un programa de computador que lo haga automáticamente. Este programa es bastante sencillo, y consiste básicamente de la lectura de las tarjetas originales y la escritura del nuevo archivo de datos en una cinta que contenga los datos de las tarjetas originales, más las imágenes de tarjetas falsas en los lugares donde sean necesarias. Si los blancos (o sus equivalentes, ceros) son códigos con significado, el usuario puede insertar tarjetas con códigos no aplicables en estos campos falsos.

Opción número dos

Incluir una instrucción especial de lectura de tarjetas en la SUBROUTINE INDEX. Puesto que la SUBROUTINE INDEX es llamada por el programa principal en el momento de la lectura de cada caso, es posible leer una tarjeta o unas tarjetas iniciales de la unidad de estudio, y en base a varios valores en uno o varios campos de datos leídos, leer o no la tarjeta adicional.

Opción número tres

Una tercera posibilidad, y probablemente la menos satisfactoria para todos, salvo los investigadores presionados por el tiempo, es la ordenación de las tarjetas de acuerdo con el número de tarjetas, y correr los programas de frecuencias y tabulaciones cruzadas sólo para tarjetas o para grupos de tarjetas que sean constantes para cada unidad de estudio. Obviamente, este método no es apropiado si se quiere tabular simultáneamente los datos que están en la parte fija como los que están en los segmentos variables de la unidad de estudio.

d) Uso de la Opción de ponderación

Un investigador puede querer ponderar las submuestras de los datos, por varias razones. Bien sea por el costo, o por la varianza diferencial esperada para las variables principales, el investigador puede tener muestreadas diferentes proporciones de la población. Además, aunque la muestra haya sido diseñada para ser autoponderada, los diferentes grados de no respuesta pueden hacer resultar desproporcionada la representación entre los diferentes estratos de la muestra. En lugar de recurrir a una sustitución, o de muestrear y duplicar casos, el sistema MINI TAB hace el proceso de ponderación relativamente fácil y poco costoso.