DEPARTAMENTO DE EVALUACION DE PROYECTOS

MANUAL DE S.P.S.S.

(STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES)

E.C.O.M.
EMPRESA NACIONAL DE COMPUTACION E INFORMATICA LTDA.
GERENCIA DE PLANIFICACION

DEPARTAMENTO DE INVESTIGACION Y DESARROLLO
JUAN JORGE MENDEZ G.

MANUAL DE S.P.S.S.
(STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES).

Casilla:14706 10/05/73 Pub. 73/49/B

I N D I C E

I N T R O D U C C I O N

I.1. Generalidades

En el análisis de datos en estudios de tipo estadístico .el especialista se encuentra, a menudo, con la dificultad de disponer de poco tiempo para el procesamiento de los datos, o de tener tal cúmulo de datos y procesos que efectuar, que esto último se vuelve casi imposible. Es así que poco a poco se ha ido popularizando el uso de computadores para tales tareas, por su rapidez y exactitud, desplazando los equipos de registro unitario, lentos y de poca flexibilidad de cálculos y procesos. La tediosa tarea de elaborar cifras y operar con los datos se desplaza entonces de el especialista al computador, permitiendo que el primero se preocupe más de planificar el estudio y analizar resultados, sin correr el riesgo de perder de vista los objetivos centrales de su estudio, debido a los largos cálculos a que se hubiera enfrentado.

Sin embargo, aún cuando los computadores procesan los datos con gran rapidez, necesitan que se les indique las operaciones a ejecutar y la secuencia en que deben ser efectuadas: necesitan ser "programados". Es por lo tanto frecuente que sea éste, motivo de desaliento del especialista estadístico, por cuanto además de saber hacerlo, programar requiere tiempo y experiencia si se quiere obtener un programa eficiente. Por lo tanto el costo del procesamiento en computador puede hacerse intolerable, máximo si para cada proceso se debe reprogramar parte de un programa o hacer uno nuevo. Por otra parte, dado que muchos estudios estadísticos usan las mismas técnicas, se observa una duplicidad de esfuerzos al confeccionarse estos programas separadamente, lo cual impide que un programa pueda ser utilizado por más de un usuario. Esto último implica que no puede repartirse el costo de un programa entre varios usuarios, ni bajarse el tiempo de programación, como sucedería si un programa confeccionado para un estudio en

particular, pudiese ser utilizado para otros.

Se hace necesario disponer de programas generalizados orientados a:

- Permitir que sea el propio especialista el que indique los procesos a efectuar mediante un lenguaje especial orientado a su problema.
- Evitar la duplicidad de trabajo, permitiendo menores costos de procesamiento.

I.2. Programa Generalizado de Procesamiento Estadístico

SPSS (Statistical Package for the Social Sciences) es un sistema integrado de programas generalizados para computador destinados principalmente al análisis de datos provenientes de estudios de tipo social, pero que son aplicables a estudios de tipo estadístico en otras disciplinas (medicina, agricultura, etc.). Inicialmente este sistema fué desarrollado en la Universidad de Stanford y actualmente es distribuido por el Centro de Investigación de la Opinión Nacional de la Universidad de Chicago.

Este sistema integrado ofrece al usuario una gran capacidad de manejo de datos y proceso estadístico de los mismos. Entre las capacidades de manejo de datos se cuenta con transformación y recodificación de variables, lo cual permite una gran flexibilidad en el trabajo ya que no se necesita alterar los datos originales para recodificar variables o definir nuevas variables a partir de otras ya existentes. La verdadera potencialidad de SPSS en cuanto al manejo de datos la irá captando el lector a medida que avance en la lectura de la presente publicación. Entre los procedimientos que SPSS puede realizar, a cada uno de los cuales corresponde un "subprograma", se puede citar el análisis de cada variable en estudio, tabulaciones de hasta diez entradas (con varios estadígrafos), correlación simple y parcial, regresión múltiple, "scaling" y análisis factorial.

El control de los subprogramas de SPSS lo tiene el usuario por medio de "tarjetas de control de SPSS' en las

cuales, con un lenguaje bastante cómodo para el especialista estadístico, especifica el manejo de datos que necesita y el o los procesos estadísticos que quiere ejecutar.

Desde el punto de vista del usuario, SPSS es una "caja negra" a la cual se le entrega datos y especificaciones, obteniéndose resultados.

I.3. Manual de uso de SPSS

Se presenta a continuación un Manual de Uso de SPSS, en el que se detallan las facilidades de SPSS, así como el "lenguaje" de control de los procesos que se desee efectuar; cualquier persona que lo lea debería poder "programar" SPSS. Consta este Manual con un anexo en el que se explica cómo especificar las características de los archivos de datos y/o resultados (tarjetas perforadas, cintas magnéticas, etc.), y un anexo con un resúmen de técnicas estadísticas para guía del usuario del Sistema. A lo largo de la descripción de los procesos, se presenta ejemplos de los listados de SPSS, para facilitar un mayor comprensión de los resultados del programa.

Este Manual cubre todas las facilidades de manejo de datos que contempla SPSS. Los procedimientos estadísticos que describe son:

- Análisis estadístico de una variable (con todos los estadígrafos que produce).
- Tabulación de variables con dos o más entradas (con todos los estadígrafos que produce).
- Correlación simple paramétrica.
- Correlación simple no paramétrica.

Las facilidades de manejo de datos y los procedimientos estadísticos que se presentan en este Manual constituyen la primera etapa de "entrega" del Sistema SPSS. Esta primera etapa contempla los procedimientos ya nombrados, debido a que son los más difundidos como primeras técnicas a usar en estudios estadísticos:

- estudiar cada variable separadamente, aislando sus características

- estudiar cada variable, aislando sus características para cada valor de otra u otras variables

- estudiar la forma como se distribuyen los individuos de una muestra con respecto a los valores de dos o más variables, obteniéndose criterios de asociación entre ellas

- estudiar la fuerza de la asociación entre variables, con técnicas paramétricas (variables numéricas, suponiendo una cierta distribución probabilística de sus valores), y técnicas no paramétricas (no suponen distribuciones probabilísticas, y se pueden emplear con variables que toman valores nominales).

I.4. Adiciones y modificaciones a este Manual

I.4.1. Adiciones

En un futuro relativamente cercano, se hará a este Manual las adiciones que corresponden a los procesos estadíticos siguientes:

- Correlación Canónica
- Correlación Parcial
- Regresión Múltiple
- "Guttman Scaling"
- Análisis Factorial

I.4.2. Modificaciones

Cualquier modificación futura a este Manual, se publicará en forma oportuna, indicando que parte es la que afecta.

I.5. Modalidad de uso de SPSS

Habiendo detectado la necesidad de disponer de un programa generalizado de procesos estadísticos, ECON ha adquirido el Sistema SPSS con el cual procesará, en principio, sus

trabajos de tipo estadístico. De acuerdo con el convenio con el organismo encargado en la Universidad de Chicago, el programa no se distribuye, y el uso de él no se cobra (solamente los recursos empleados: tiempo de computador, papel, tarjetas, etc) dado que una de las finalidades de ECOM es prestar servicio a Instituciones del Sector Público y Area Social.

I.6. Otro programa de uso generalizado

En el futuro, ECOM contará con otro programa generalizado, OSIRIS (Organized Set of Integrated Routines for Investigations with Statistics), el cual es un potentísimo programa de procesos estadísticos que cuenta con mayores posibilidades de manejo de datos y con procedimientos estadísticos avanzados de gran variedad. Ambos programas tienen una interfase de comunicación del uno con el otro, por lo cual se puede usar SPSS u OSIRIS para diferentes partes de un estudio, de acuerdo con las distintas complejidades que pudieran tener. Para la realidad Nacional parecía más conveniente implementar primero SPSS, el cual contiene los procedimientos básicos de análisis estadístico. Sin embargo se prevé que en un futuro próximo el procesamiento estadístico de datos hará también uso de las técnicas estadísticas más sofisticadas que contiene OSIRIS.

1. ESTRUCTURA DEL SISTEMA SPSS

El SPSS es un sistema integrado para el análisis estadístico de datos. Está estructurado en base a subprogramas, es decir, cada proceso estadístico o de manejo de información, es realizado por un subprograma específico. Reúne, este sistema, amplias facilidades para el manejo de la información y la mayoría de las técnicas estadísticas más usuales en estudios de tipo social.

Realizar un proceso usando este sistema requiere en primer lugar, indicarle al computador que acciones tomar para usar el sistema. Esto se hace mediante una serie de instrucciones, en un lenguaje especial (Job Control Language) que están dirigidas a los programas que controlan la operación interna del computador. En nuestro caso corresponden a lo que se llama Sistema Operativo (O.S.). En segundo lugar se debe indicar al sistema, ya listo para ser usado, que proceso de los que ofrece SPSS quiere realizarse y cuáles son las características de los datos sobre los que se efectuará el proceso. Esto se hace mediante tarjetas de control del Sistema. Previamente el usuario debe codificar los datos y almacenarlos en algún dispositivo desde el cual puedan ser leídos por el computador. Este dispositivo puede ser tarjetas perforadas, cintas o discos magnéticos.

Así un programa de SPSS está formado por:

a) Las tarjetas de control del O.S.

b) Las tarjetas del SPSS.

c) Las 'tarjetas' de datos.

1.1. Tarjetas de Control del O.S.

Estas tarjetas se perforan siempre a partir de la primera columna y pueden ocupar hasta la columna 71 inclusive. Además siempre empiezan por slash '/'.

Las tarjetas son:

```
1                16
//SPSSH PROC
//STEP EXEC    PGM=SPSS,PARM=50000
//STEPLIB DD   DSN=SYS2.SPSSHLIB,DISP=SHR,UNIT=2314,VOL=SER='ET20B
//FT01F001 DD  UNIT=SYSDA,SPACE=(800,(100,50))
//FT02F001 DD  UNIT=SYSDA,SPACE=(2012,(200,200))
//FT06F001 DD  SYSOUT=A
//SYSABEND DD  SYSOUT=A
//FT05F001 DD  DDNAME=SYSIN
//ENDED PEND
//A EXEC SPSSH
//SYSIN DD *
.............  }
.............  }
.............  } Tarjetas del SPSS
.............  }
.............  }
.............  } Tarjetas de Datos
.............  }
.............
/*
```

Nota:

PARM=50000 Corresponde al espacio asignado en memoria. Este parámetro (=SPACIO) es modificable por el usuario, pero el valor mayor que puede tener para partición de 180 Kb es 50.000 bytes.

Las tarjetas de control pueden sufrir modificaciones posteriormente. Estas serán comunicadas en forma oportuna a los usuarios.

1.2. Tarjetas del Sistema.Primera Parte

Las tarjetas del sistema pueden dividirse en, tarjetas de control, tarjetas de descripción de archivos, tarjetas de control de procesos estadísticos, tarjetas de modificación de datos y tarjetas de manejo de archivos.

Todas ellas tienen en común el estar compuestas de dos campos:

a) Campo de Control: que ocupa las columnas 1 a 15 inclusive y que contiene una palabra clave de identificación.

b) Campo de Especificación: que ocupa las columnas 16 a 80 inclusive y cuyo contenido varía de acuerdo a la tarjeta.

Ejemplo:

Col.1	Col.16	Col.80
VARIABLES LIST	VARBLE1,VARBLE2,VARBLE3,.....	
palabra clave en el campo de control	lista de variables declaradas en el campo de especificación	

1.2.1. Tarjetas de Control del Sistema SPSS

1.2.1.1. Tarjeta RUN NAME

Sirve para identificar el proceso para el usuario.
Su formato es:

```
1          16
RUN NAME   rótulo
```

El rótulo puede contener hasta 64 caracteres, incluyendo cualquier carácter válido de una perforadora IBM-029. Este rótulo se imprime al comienzo de cada página de listado.

Esta tarjeta es opcional.

1.2.1.2. Tarjeta READ INPUT DATA

Esta tarjeta no tiene campo de especificación y su función es informar al sistema que a continuación deben leerse los datos. Es siempre requerida cuando se leen los datos desde un archivo que no ha sido generado por el sistema. Su formato es:

```
1
READ INPUT DATA
```

1.2.1.3. Tarjeta FINISH

Esta tarjeta tampoco tiene campo de especificación y su función es indicar al sistema que se termine el proceso.

Su formato es:

```
1
FINISH
```

La ubicación de estas tarjetas se muestra en el siguiente ejemplo:

```
1                16
//SPSSH PROC
:
:
//SYSIN DD *
RUN NAME       EJEMPLO PARA INDICAR POSICION TARJETAS DE CONTROL
.............
.............  } Otras tarjetas de SPSS
.............
READ INPUT DATA
.............
.............  } Tarjetas de Datos
.............
FINISH
/*
```

1.2.2. Tarjeta de Definición de Archivos

Un 'caso estadístico', es decir, un objeto o individuo para el cual se han registrado ciertas mediciones, es la unidad básica de análisis estadístico. Estas mediciones se denominan variables estadísticas y el conjunto de casos se denomina archivo (FILE).

Así, por ejemplo, en una empresa podría interesar investigar las características y relaciones entre insumos, producción y ventas semanales. El caso estadístico sería una semana cualquiera para la cual se conociera el valor de los insumos, producción y ventas de la misma. El archivo estaría constituído por las semanas correspondientes al período que se quiere investigar.

Desde el punto de vista del sistema, existen dos tipos de archivos, los que proporciona el usuario y lo que genera el sistema a partir de datos o casos del usuario. Las diferencias son básicamente computacionales y son dos las principales: una es que los archivos que genera el sistema se pueden almacenar sólo en cintas o discos magnéticos. La otra, es que los archivos del sistema guardan la información preprocesada, incluyendo una descripción del archivo (nombres de variables, numero de casos, tipos de variables, etc). Esta segunda diferencia, impide que estos archivos puedan ser leídos por programas ajenos al sistema.

La ventaja de los archivos generados por el sistema es su fácil y rápido manejo por el sistema mismo, lo que resulta muy conveniente en el caso en que sobre un mismo conjunto de datos vayan a realizarse varios procesos sucesivos por el sistema.

1.2.2.1. Tarjeta FILE NAME

Esta tarjeta permite asignarle un nombre al archivo del usuario. Su formato es:

1 16
FILE NAME nombre (rótulo opcional)

Nombre: hasta 8 caracteres, el primero alfabético.
Rótulo opcional: hasta 62 caracteres, sin caracteres especiales (,+-/*()'=.)

Esta tarjeta es obligatoria sólo cuando el usuario quiere guardar su archivo como archivo del sistema (ver punto 4.2.).

1.2.2.2. Tarjeta VARIABLE LIST

Esta tarjeta sirve para asignar un nombre a cada una de las variables. Su formato es:

1 16
VARIABLE LIST lista de nombres de variables

Los nombres de variables pueden tener hasta 8 caracteres alfanuméricos, el primero siempre alfabético. Deben ir separados por una coma y/o uno o más blancos. Deben ser únicos en términos de los nombres de las otras variables. El orden en que se declaren debe ser el mismo que tienen las variables en el archivo.

Si no alcanzara una tarjeta para la especificación de las variables, se puede continuar en tarjetas sucesivas dejando en blanco el campo de control de las tarjetas de continuación (cols. 1 a 15). No se puede romper el nombre de una variable en dos tarjetas.

No se pueden declarar más de 500 variables.

Ejemplo:

```
1               16
VARIABLE LIST   NOMBRE,APELLIDO,DIREC,SEXO,EDAD,INGRESO
                EDUC,ESTCIV,HIJOS
```

Cuando el archivo tiene muchas variables, se les puede asociar nombres mediante secuencia de números; con esto se evita tener que asignar nombres distintos a demasiadas variables. Esto se hace mediante la siguiente especificación:

VARxxx TO VARyyy

Los xxx e yyy son números de tres dígitos tales que xxx es menor que yyy. De esta forma se aignan nombres a yyy-xxx+1 variables en forma correlativa.

Ejemplo:

```
1                16
VARIABLE LIST  MARCA,MODELO,VAR001 TO VAR025,COLOR
```

En total se declaran 28 variables, MARCA,MODELO, VAR001,VAR002,.....VAR025,COLOR.

1.2.2.3. Tarjeta INPUT MEDIUM

Esta tarjeta informa al sistema sobre el dispositivo en que reside el archivo del usuario. Este dispositivo puede ser tarjetas perforadas, cintas o discos magnéticos (ver punto 4.1.), por el momento asumiremos que el archivo está en tarjetas. El formato es:

```
1              16
INPUT MEDIUM   CARD
```

1.2.2.4. Tarjeta INPUT FORMAT

Los datos del archivo, en tarjetas, pueden estar codificados en dos tipos de formatos.

1.2.2.4.1. Formato de Campo Libre

Los valores de las variables en las tarjetas deben ir separados por una coma y/o blancos y los valores de las variables alfanuméricas deben ir entre comillas.

La especificación es:

```
1              16
INPUT FORMAT   FREEFIELD
```

1.2.2.4.2. Formato Fijo

En este formato los valores de cada una de las variables deben ocupar las mismas columnas y la misma tarjeta dentro de cada caso, por ejemplo : las columnas 12 a 16 de la tarjeta número 3 de cada caso. Un valor no puede ser perforado en dos tarjetas, debe estar contenido totalmente en una sola tarjeta.

La especificación es:

```
1                16
INPUT FORMAT     FIXED (lista de formatos)
```

Lista de Formatos: esta lista indica el tipo y localización de cada variable dentro del caso. Esto se hace a través de cuatro elementos de de formato que son:

1) 'A', para variables alfanuméricas

Estructura: nAw — n variables contiguas que ocupan w columnas cada una y que son alfanuméricas. Si n = 1 puede omitirse.

2) 'F', para variables numéricas

Estructura: nFw.d — n variables contiguas que ocupan w columnas cada una (incluyendo el signo y el punto decimal si se ha perforado), con d dígitos o columnas a la derecha del punto decimal. Si no se ha perforado el punto decimal en todas las tarjetas, debe usar el d, para indicar el N°de decimales. Si n=1 puede omitirse.

Ejemplo: -7.32 con F5.2 lee - 7.32
-7.32 con F5.0 lee - 7.32(prima el punto perforado)
-732 con F4.2 lee - 7.32

3) 'X', para saltar columnas.

Estructura: mX — deja de leer (o se salta) m columnas (m no puede omitirse).

4) '/' , para saltar tarjetas

Estructura : / indica al sistema que debe continuar leyendo en la siguiente tarjeta.

Ejemplo:

INPUT FORMAT FIXED(10X,5F2.0,3X,5A1,2X,3F5.1/10X,22A2,5X,F5.0)

10X .- Se salta diez columnas
5F2.0 .- Lee cinco variables numéricas de dos columnas cada una
3X .- Se salta 3 columnas
5A1 .- Lee cinco variables alfanuméricas de una columna cada una
2X .- Se salta 2 columnas
3F5.1 .- Lee tres variables numéricas de cinco columnas, con un decimal
/ .- Va a la segunda tarjeta
10X .- Se salta diez columnas
22A2 .- Lee veintidos variables alfanuméricas de dos columnas cada una
5X .- Se salta cinco columnas
F5.0 .- Lee una variable numérica de cinco columnas.
( ) .- Repite lo anterior para cada caso

1.2.2.5. Tarjeta # OF CASES

Esta tarjeta informa al sistema el número de casos (no de tarjetas), que contiene el archivo.

Su formato es:

1	16
# OF CASES	NUMERO DE CASOS

1.2.2.6. Tarjeta PRINT FORMATS

Sirve para indicar el número de decimales con que el usuario desea que se impriman las variables numéricas en los listados de los procesos estadísticos. El máximo de decimales es cinco y si no se indican asume dos. Para variables numéricas es opcional.

La otra función es indicar, para los efectos de impresión, cuando una variable es alfanumérica. En este caso es obligatoria.

Su formato es:

```
1                16
PRINT FORMATS    Lista de variables (Valor) / lista
                 de variables (Valor)
```

La lista de variables puede contener cualquier variable declarada en la tarjeta VARIABLE LIST, cuando contiene más una variable, éstas deben ir separadas por una coma y/o uno o más blancos. Cuando se declara el mismo valor para un conjunto de variables adyacentes en la tarjeta VARIABLE LIST, puede hacerse más fácilmente nombrando la primera variable, poniendo la palabra clave TO y nombrando la última variable del conjunto.

Ejemplo:

```
1                16
VARIABLE LIST    A,B,C,DELTA, K  , Z , VAR ØØ1 TO VARØ22,

PRINT FORMATS    A TO DELTA (A)/ K (3) / Z TO VARØ17 (4)
                 VARØ18 TO VARØ22 (Ø)
```

Este tipo de especificación para variables adyacentes es bastante usado en el sistema. En adelante nos referiremos a el como 'especificación del tipo VARA TO VARZ'.

El valor, que sigue a la lista de variables, debe ir entre paréntesis y puede ser:

- para variables numéricas: Ø, 1, 3, 4, 5
- para variables alfanuméricas: A

1.2.2.7. Posición de las Tarjetas dentro de un Programa

```
1               16
//SPSSH PROC
.............
.............
.............
//SYSIN DD*
RUN NAME        EJEMPLO N.2 INDICANDO ORDEN TARJETAS DESCRIP.DE ARCH.
FILE NAME       PRUEBA DATOS PARA EJEMPLO N2
VARIABLE LIST   CLAVE,EDAD,SEXO,DIAGNOS,VARØØ1 TO VARØ3Ø,RESUL
INPUT MEDIUM    CARD
INPUT FORMAT    FIXED(A4,F3.Ø,1ØX,F1.0,5X,A8/3ØF2.0,2X,A8)
Ø OF CASES      1ØØ
PRINT FORMATS   VARØØ1 TO VARØ3Ø (0)/CLAVE,DIAGNOS,RESUL(A)
.............   }
.............   } Tarjetas de selección del proceso estadístico
.............   }
READ INPUT DATA
.............   }
.............   } Datos (Archivo del usuario)
.............   }
FINISH
/*
```

1.2.3. Tarjetas de Control de Procesos Estadísticos

Los procesos estadísticos están estructurados como subprogramas. Por esto es necesario seleccionarlos, indicarles sobre cuales variables realizar el proceso, qué estadísticos se desea que calcule y otras informaciones,(como tipo de impresión, por ejemplo) que son especificas de cada subprograma estadístico.

1.2.3.1. Tarjeta de Procedimiento.

Mediante esta tarjeta se selecciona y se indican las variables que van a intervenir en el mismo. En general, en su campo de control va una palabra clave que identifica el proceso y en su campo de especificación se incluyen las variables.

Su formato está dado por el proceso requerido y en cada caso indica al sistema que debe realizarse el proceso seleccionado, sobre las variables especificadas.

Ejemplo:

```
1                16
CONDESCRIPTIVE   EDAD, VARØØ1 TO VARØØ9
```

Se realiza el proceso CONDESCRIPTIVE sobre las variables EDAD, VARØØ1, VARØØ2, ..., VARØØ9 (que han sido previamente declaradas en la tarjeta VARIABLE LIST).

1.2.3.2. Tarjetas de Estadísticos

Con esta tarjeta se indican los estadísticos que el usuario desea del proceso. Mediante un código numérico, el usuario selecciona los estadísticos deseados de los disponibles en el proceso.

Esta tarjeta puede ser opcional u obligatoria de acuerdo al proceso.

Su formato es:

```
1           16
STATISTICS  { lista de estadísticos }
            { ALL                   }
```

La lista de estadísticos varía según los procesos. En los que veremos en el próximo capítulo, la lista de estadísticos es muy similar y tiene en común, los siguientes estadísticos*:

Código	Estadístico
1	Media
2	Error Standard (de la media)
5	Desviación Standard
6	Varianza
7	Kurtosis
8	Asimetría (Skewness)
9	Rango
10	Mínimo
11	Máximo

* Ver Anexo Estadístico

La clave ALL indica que se calculen e imprimen todos los estadísticos disponibles en el proceso.

1.2.3.3. Tarjeta de Opciones

Esta tarjeta proporciona información adicional para controlar los cálculos que han sido solicitados en la tarjeta de procedimiento. Su formato general es:

```
1        16
OPTIONS  lista de opciones
```

La lista de opciones es específica de cada procedimiento y se irá explicando a medida que se vayan necesitando.

Esta tarjeta se ubica a continuación de la tarjeta de procedimiento, cuando se especifica. Es opcional.

Ejemplo:

```
1          16
//SPSSH PROC
:
:
:
//SYSIN DD *
RUN NAME        PRUEBA PROGRAMA COMPLETO
FILE NAME       ARCH PRUEBA PILOTO
VARIABLE LIST   VARA,VAR00001 TO VAR00025,NUM,REG
INPUT MEDIUM    CARD
INPUT FORMAT    FREEFIELD
# OF CASES      220
PRINT FORMATS   VARA TO REG (2)
*  CONDESCRIPTIVE ALL
*  OPTIONS        1,2
*  STATISTICS     ALL
READ INPUT DATA
:
:   Datos
:
:
FINISH
/*
```

---

* Se verán en el próximo capítulo

2. PROCEDIMIENTOS ESTADISTICOS, I

Los cuatro subprogramas estadísticos que se explican a continuación, calculan estadísticas descriptivas para una variable. El primero de ellos es el único que trabaja con variables contínuas (no categorizadas, es decir toman un número muy grande de valores distintos), los otros tres trabajan con variables numéricas discretas y variables alfanuméricas. Estas restricciones no presentan inconvenientes graves ya que es posible transformar, por medio del sistema, variables continuas a discretas, alfanuméricas a numéricas y así aprovechar al máximo las ventajas de uno u otro subprograma.

2.1 Estadísticos Descriptivos para Variables Contínuas

Subprograma CONDESCRIPTIVE

El listado de este programa está dado por los estadísticos que se soliciten en la tarjeta de estadísticos, por lo que en este subprograma esa especificación es obligatoria.

2.1.1. Tarjeta de Procedimiento

1	16
CONDESCRIPTIVE	lista de variables
	ALL

La lista de variables se especifica en la misma forma que en el caso de la tarjeta PRINT FORMATS. Es decir los nombres de las variables deben ir separados por una coma y/o uno o mas blancos y se puede hacer mención a variables adyacentes en la declaración de la tarjeta VARIABLE LIST, mediante la especificación VARA TO VARZ p. ej. VARA TO REG del ejemplo anterior.

La palabra clave ALL significa que se incluyen todas las variables declaradas en el archivo.

2.1.2. Lista de estadísticos

Tiene sólo los estadísticos comunes mencionados en el punto 1.2.3.2. La tarjeta ya se ha descrito en ese mismo punto.

2.1.3. Opciones

Las opciones de este subprograma se verán en el punto 3.2.1.

2.1.4. Limitaciones del subprograma

1. No mas de 250 nombres de variables pueden ser especificados en una tarjeta CONDESCRIPTIVE (Las especificaciones de la forma VARA TO VARZ cuentan sólo como 3)

2.2. Estadísticos Descriptivos para Variables Discretas.Histogramas

2.2.1 Subprograma CODEBOOK

Este subprograma confecciona tablas de frecuencia, frecuencia relativa y frecuencia acumulada.Opcionalmente el usuario puede pedir histogramas.(Ver ejemplo punto 3.2.1)

Puede operar con variables numéricas o alfanuméricas, sin embargo los estadísticos son calculados sólo para variables numéricas.

El uso mas eficiente de este programa se logra cuando las variables están agrupadas en no más de 20 categorías.

2.2.1.1. Tarjeta de Procedimiento

```
1
CODEBOOK { 16
         { lista de variables o
         { ALL
```

La lista tiene la misma estructura que en el programa CONDESCRIPTIVE.

2.2.1.2. Lista de Estadísticos

Tiene los estadísticos comunes mencionados en el punto 1.2.3.2., mas otros dos que son:

FILE DATOS (CREATION DATE = 03/13/73) FICTICIOS DE EJEMPLO

VARIABLE EDUC

VALUE LABEL	VALUE	ABSOLUTE FREQUENCY	RELATIVE FREQUENCY (PERCENT)	ADJUSTED FREQUENCY (PERCENT)	CUMULATIVE ADJ FREQ (PERCENT)
	E	23	38.3	38.3	38.3
	D	19	31.7	31.7	70.0
	T	10	16.7	16.7	86.7
	U	8	13.3	13.3	100.0
	TOTAL	60	100.0	100.0	100.0

VALID OBSERVATIONS = 60
MISSING OBSERVATIONS = 0

Código	Estadístico
3	Mediana
4	Moda

2.2.1.3. Lista de Opciones:

Por el momento sólo se especifican dos, las demás se especifican en el punto 3.2.2.

Opción 4 indica que se impriman histogramas de cada una de las variables especificadas en la tarjeta CODEBOOK.

Opción 5 indican que se impriman los histogramas y que no se impriman las tablas de frecuencia.

Si no se pone la tarjeta de opciones, se imprimen las tablas de frecuencia y no se imprimen histogramas.

2.2.1.4. Limitaciones del subprograma

1.- Análoga a la limitación de CONDESCRIPTIVE.

2.- El número máximo de variables que pueden ser procesadas, es función del número de categorías en la variable que tiene mas. Esta dependencia está dada por:

$$MAXVAR = \frac{(ESPACIO/4) - (8 \times NCAT) - 15}{4+(2 \times NCAT)}$$

donde MAXVARS es el número máximo de variables.

NCAT es el número de categorías en la variable que tiene más.

ESPACIO: es la cantidad de memoria asignada en el parámetro PARM de la tarjeta de control del O.S. correspondiente.

Por ejemplo con PARM=50.000 y 20 categorías como tope, se tiene que el número máximo de variables, MAXVARS, es aproximadamente 280.

2.2.2. Subprograma MARGINALS

Básicamente ejecuta los mismos cálculos que CODEBOOK, frecuencia absoluta, relativa y acumulada. No produ-

ce histogramas pero puede trabajar con variables numéricas o alfanuméricas que tengan un gran número de categorías.

2.2.2.1 Tarjeta de Procedimientos

```
1                16
MARGINALS  { lista de variables o
           { ALL
```

La lista de variables tiene la misma estructura que la de los programas anteriores.

2.2.2.2 Lista de Estadísticos

Es la misma de CODEBOOK

2.2.2.3 Lista de Opciones

Por el momento se dan solo dos.

Opción 3 Suprime la impresión de frecuencia acumulada.

Opción 5 Suprime la impresión de tablas de frecuencia, imprimiendo solo los estadísticos.

Si no se pone la tarjeta de opciones se imprime la frecuencia acumulada y las otras tablas.

2.2.2.4 Limitaciones del subprograma

1. La misma que en los casos anteriores.
2. El número máximo de variables que puede ser procesado está dado por:

$$MAXVAR = \frac{(ESPACIO/4)-(3xNCATE)}{2+(2*NCATE)}$$

Donde MAXVAR es el número máximo de variables

NCATE el número de categorías en la variable que tenga mas.

ESPACIO es la memoria asignada en el parámetro PARM.

2.2.3 Subprograma FASTMARG

Este subprograma tiene básicamente los mismos listados que CODEBOOK (excepto histogramas). La principal diferencia es que sólo trabaja con variables numéricas, su ventaja es que es aproximadamente el doble mas rápido que los anteriores.

2.2.3.1 Tarjeta de Procedimiento

```
1                16
FASTMARC         lista de variables (mínimo,máximo)/
                 lista de variables (mínimo, máximo)
```

En este caso además de la lista de variables (similar a la de los casos anteriores) debe indicarse el valor mínimo y máximo de las variables *

Ej;

```
1                16
FASTMARC         EDAD (0,120)/VARØØ1 TO VARØ25 (0,100)/
                 VARØ26, EDUC, EST (0,5)/VARØ27 TO VARØ51(0,200)
```

2.2.3.2 Lista de Estadísticos

Los mismos de CODEBOOK

2.2.3.2 Lista de opciones:

Se dan sólo dos por el momento:

Opción 4 Produce que los resultados salgan a unidad de cintas o discos de tal manera que se puedan sacar copias posteriormente. Para esto debe prepararse una tarjeta de control //FT09F001 (Ver Anexo E) que se pone a continuación de //A EXEC SPSSH

Opción 5 Suprime la impresión de distribuciones de frecuencias, imprimiendo solo los estadísticos.

2.2.3.3 Limitaciones del Subprograma

1. El espacio que FASTMARG necesita para confeccionar las tablas está dado por:

   Espacio= 4xSUMVAL

   SUMVAL: es la suma del número de categorías en cada variable.

* Conviene que los valores de las variables sean adyacentes con el fin de ahorrar espacio, por ejemplo: si los valores 0,1,2,3, 9, el mínimo en 0 y el máximo 9 y el subprograma reserva espacio para 10 categorías cuando en realidad hay sólo 5.

Ej. Considerando el ejemplo anterior tendríamos

EDAD	121
VAR001 TO VAR025	25x101
VAL026, EDUC, EST	3x6
VAR027 TO VAR050	25x201
	7.689

Espacio = 4x7.689 = 30.756

Si el ESPACIO asignado en el parámetro PARM es mayor que de 31.000, no se producen problemas.

3. TARJETAS DEL SISTEMA
SEGUNDA PARTE

3.1. Tarjetas Adicionales de Descripción de Archivos

El sistema ofrece la posibilidad de asignar rótulos explicativos a las variables y a las categorías de las variables con el fin de proporcionar listados ampliamente documentados. Además permite dar un tratamiento estadístico diferenciado a aquellos valores que corresponden a valores faltantes. (Ej.: no contesta).

3.1.1. Rótulos de Variables

SPSS permite asociar a cada variable un rótulo de hasta 40 caracteres, que se imprime en todos los listados en que aparezca esa variable. Esto se hace mediante la instrucción.

```
1          16
VAR LABELS variable, rótulo/variable,rótulo/...
           /variable,rótulo.....
```

Los rótulos pueden contener cualquier carácter válido en la perforadora IBM-029 excepto el slach '/', la coma y los paréntesis '()'.

3.1.2. Rótulos de Valores de Variables

Es también posible asociar un rótulo a cada valor de cualquiera o de todas las variables de un archivo. Estos rótulos pueden ser extremadamente útiles para documentar los resultados de distribuciones de frecuencias simples y tabulaciones. La especificación es:

```
1                16
VALUE LABEL   lista de variables  (valor 1)rótulo 1..
              ...(valor n)rótulo n /

              lista de variables  (valor 1)rótulo 1
              (valor 2) rótulo 2
```

La lista de variables se especifica de la misma forma que en el caso de la tarjeta PRINT FORMATS.

Los rótulos son de hasta 16 caracteres y pueden contener cualquier carácter válido excepto el slach y los paréntesis.

3.1.3. Valores Faltantes

Es frecuente que en una muestra estadística algunos casos no tengan la información completa. Por ejemplo, en un estudio socioeconómico, en algunos de los casos no hay información acerca del ingreso. Para no excluir estos casos del estudio y para que, si se incluyen, no distorsionen los resultados, el usuario puede poner una clave numérica o alfanumérica, para indicar que no se tiene información acerca del ingreso en esos casos particulares.

Para informar al sistema acerca de estas claves, se usa la tarjeta MISSING VALUES.

Ejemplo:

```
1              16
MISSING VALUES  INGRESO (-1)
```

En este ejemplo se indica al sistema que el valor de -1 para la variable INGRESO indica que no hay información.

El formato general es:

```
1              16
MISSING VALUES  {lista de variables}(lista de valore
                {lista de variables}(lista de valore
```

La lista de variables acepta especificaciones del tipo VARA TO VARZ.

En la lista de valores, estos deben ir separados por una coma y/o uno o más blancos. Los valores alfanuméricos deben ir entre comillas. No pueden especificarse más de tres valores como valores faltantes por variable.

3.2. Opciones Adicionales para Procesos Estadísticos

En relación a las tarjetas de definición de archivos, ya descritos, existen, para cada subprograma de procesos estadísticos visto anteriormente, opciones referentes a la impresión o no impresión de rótulos y al tratamiento de los valores faltantes.

3.2.1. CONDESCRIPTIVE

Opción 1

El sistema ignora los valores declarados faltantes e incluye todos los valores de las variables en el cálculo.

Opción 2

Suprime la impresión de rótulos de variables. Esta opción aumenta la rapidez del subprograma.

Si no se especifican estas opciones el subprograma asume:

a) Se consideran las especificaciones de valores faltantes y no se incluyen estos valores en los cálculos.

b) Los rótulos de variables que han sido declarados se imprimen.

3.2.2. CODEBOOK

Opción 1

El sistema ignora los valores declarados faltantes e incluye todos los valores en los cálculos.

Opción 2

Suprime la impresión de rótulos. Esta opción aumenta la rapidez del subprograma. (Si el usuario no desea imprimir rótulos de valores y no desea

EJEMPLOS 03/13/73 PAGE 8

FILE DATOS (CREATION DATE = 03/13/73) FICTICIOS DE EJEMPLO

VARIABLE EDAD

VALUE LABEL	VALUE	ABSOLUTE FREQUENCY	RELATIVE FREQUENCY (PERCENT)	ADJUSTED FREQUENCY (PERCENT)	CUMULATIVE ADJ FREQ (PERCENT)
MENOS DE 25	1.00	27	45.0	45.0	[illegible]
ENTRE 25 Y 30	2.00	11	18.3	18.3	63.3
ENTRE 30 Y 35	3.00	11	18.3	18.3	81.7
ENTRE 35Y 40	4.00	5	8.3	8.3	90.0
ENTRE 40 Y 50	5.00	2	3.3	3.3	93.3
ENTRE 45 Y 50	6.00	1	1.7	1.7	95.0
MAS DE 50	7.00	3	5.0	5.0	100.0
	TOTAL	60	100.0	100.0	100.0

EJEMPLOS

FILE DATOS (CREATION DATE = 03/13/73) FICTICIOS DE EJEMPLO

VARIABLE EDAD

CODE			
1.00	MENOS DE 25	(27)	45.0 PCT
2.00	ENTRE 25 Y 30	(11)	18.3 PCT
3.00	ENTRE 30 Y 35	(11)	18.3 PCT
4.00	ENTRE 35Y 40	(5)	8.3 PCT
5.00	ENTRE 40 Y 50	(2)	3.3 PCT
6.00	ENTRE 45 Y 50	(1)	1.7 PCT
7.00	MAS DE 50	(3)	5.0 PCT

0 5 10 15 20 25 30 35 40 45 50

FREQUENCY

STATISTICS

MEAN	2.317	STD ERROR	0.213	MEDIAN	1.773
MODE	1.000	STD DEV	1.652	VARIANCE	2.729
KURTOSIS	1.294	SKEWNESS	1.373	RANGE	6.000
MINIMUM	1.000	MAXIMUM	7.000		

VALID OBSERVATIONS = 60
MISSING OBSERVATIONS = 0

histogramas, es más conveniente que use el subprograma MARGINALS que en ese caso es más eficiente).

Si no se especifican opciones el subprograma asume:

a) Se consideran las especificaciones de valores faltantes y no se incluyen estos valores en los cálculos.

b) Los rótulos que han sido declarados se imprimen.

3.2.3. MARGINALS

Opción 1

id. CONDESCRIPTIVE

Opción 2

id. CONDESCRIPTIVE

Opción 6

Los valores faltantes son incluídos en las tablas pero no en el cálculo de los estadísticos.

Si no se especifican las opciones, se excluyen los valores faltantes, se imprimen rótulos y no tiene efecto la opción 6.

3.2.4. FASTMARG

Opción 1

Se ignoran las declaraciones de valores faltantes y se consideran todos los valores dentro del rango especificado.

Opción 2

Suprime los rótulos de valores en los listados. Esta opción puede aumentar considerablemente la capacidad del programa ya que en otro caso el espacio requerido es:

Espacio $4 \times$ SUMVAL $\times 20 \times$ MAXROT

donde MAXROT es el máximo número de categorías de las variables. Si se usa la opción 2 o 5 MAXROT=0.

Opción 3

Todas las tablas y estadísticos son impresos en la porción izquierda de la hoja en un formato de aproximadamente 20x27 cms. Se elimina el nombre del programa, el nombre del archivo y todos los rótulos de los listados

Si no se especifican estas opciones; no se incluyen los valores faltantes, se imprimen los rótulos y se imprimen los resultados en formato normal.

4. ARCHIVOS EN DISPOSITIVOS MAGNETICO

Existen dos tipos de archivos:

- archivos del usuario
- archivo del sistema

Los datos originales del usuario, que se ingresan al sistema, conforman lo que se llama archivo del usuario. Estos datos pueden estar perforados en tarjetas o grabados (mediante algún otro programa especial) en cintas o discos magnéticos.

SPSS puede grabar los datos provenientes del archivo del usuario, en forma especial, dando orígen a un archivo del sistema. En dicho archivo, para cada variable, se incluye el nombre de la variable, los rótulos asociados, los valores faltantes, el tipo de variable y el formato de salida, además de otra información. Esto permite que a partir de esa grabación no sea necesario describir el archivo nuevamente en procesos posteriores.

Para leer y/o graber estos archivos aparecen nuevas instrucciones y nuevos parámetros para instrucciones ya conocidas. Además es necesario preparar también nuevas tarjetas de control del O.S. para cada tipo de entrada o salida. (Ver Anexo E).

4.1. Archivos del Usuario en Dispositivos Magnéticos

Desde el punto de vista de las tarjetas del sistema SPSS, prácticamente no cambia en nada la estructura del proceso de los archivos en dispositivos magnéticos. Las únicas diferencias están en las tarjetas INPUT MEDIUM y # OF CASES

4.1.2. Tarjeta INPUT MEDIUM

Esta tarjeta cuenta ahora con dos nuevos parámetros, estos son:

```
1                16
INPUT MEDIUM TAPE      para leer desde cintas
INPUT MEDIUM DISK      para leer desde discos
```

4.1.3. Tarjeta # OF CASES

Cuando los archivos residen en dispositivos magnéticos, pueda darse que no se sepa exactamente el número de casos del archivo. Para obviar este problema existe un tipo especial de especificación, sólo para estos casos, en que basta indicar el número estimado de casos. Su formato es:

```
1                16
# OF CASES   ESTIMATED (número estimado de casos)
```

Ej.: El archivo del usuario ha sido grabado en una cinta con el nombre PRUEBA, en registro de largo 60, con 40 registros por bloque, con labels = standard, siendo el primer archivo dentro de la cinta '804'

```
1                16
//SPSSH PROC
//STEP EXEC PGM=SPSS,PARM=50.000
:
:
//A EXEC SPSSH
//FT08F001 DD  DSN=PRUEBA,UNIT=2400,LABEL=(1,SL),DISP=OLD,
//             DCB=(RECFM=FB,LRECL=60,BLKSIZE=2400),VOL=SER=804
//SYSIN DD*
RUN NAME        LECTURA DESDE CINTA MAGNETICA
VARIABLE LIST   VAR001 TO VAR020
INPUT MEDIUM    TAPE
INPUT FORMAT    FIXED(20F2.0)
# OF CASES      ESTIMATED 1000
MISSING VALUES VAR001 TO VAR015 (-1,0)/VAR017 TO VAR029 (98,99)
```

```
CONDESCRIPTIVE VAR017 TO VAR025
STATISTICS     ALL
READ INPUT DATA
CODEBOOK       VAR001 TO VAR005
STATISTICS     3,4
OPTIONS        5
MARGINALS      VAR010
STATISTICS     ALL
FINISH
/*
```

4.2. Archivos del Sistema

Cuando sobre un mismo set de datos van realizarse distintos procesos en varias etapas, el usuario puede grabar su archivo como archivo del sistema. Esto permite leer la información directamente desde tal archivo en los procesos posteriores. Las ventajas son significativas; se requieren menos tarjetas de control del SPSS y hay un ahorro de tiempo considerable sobre todo cuando se trabaja con grandes cantidades de datos.

Su utilidad se apreciará aún más cuando se analicen los procesos de transformación y recodificación de variables y el capítulo de manejo de archivos del sistema.

4.2.1. Grabación de un Archivo del Sistema

Este proceso de grabación requiere de una tarjeta de control del O.S. (ver Anexo E) y una tarjeta de control del SPSS. Esta última va ubicada inmediatamente antes de la tarjeta FINISH. Su formato es:

```
1
SAVE FILE
```

Es obligatorio que el usuario asigne un nombre al archivo del sistema que va a generar, mediante la tarjeta FILE NAME. Este nombre será usado luego en los procesos de lectura de estos archivos.

La grabación del archivo se hace posteriormente a todos los procesos estadísticos y los procesos de

transformación y recodificación de variables.

4.2.2. Lectura de Archivos del Sistema

Una vez grabado el archivo del usuario como archivo del sistema, para volver a leerlo en un proceso posterior basta usar una tarjeta del sistema. Además de las tarjetas de control del O.S.

El formato de esta tarjeta es:

1	16
GET FILE	'nombre del archivo especificado en la tarjeta FILE NAME en el proceso de grabación'

Esta tarjeta sigue a la tarjeta RUN NAME.

Ejemplo N°1.-

```
1                16
//SPSSH PROC
//
//
//A EXEC SPSSH
//FT04F001 DD  DSN=ARS,UNIT=2400,LABEL=(1,SL),DISP=(NEW,KEEP),
//            DCB=BLKSIZE=4000,VOL=SER=809
//
//SYSIN DD*
RUN NAME        GRABACION DE ARCHIVO DEL SISTEMA
FILE NAME       CUENTAS
VARIABLE LIST   IDENT,NCUEN,VAR014 TO VAR041
INPUT MEDIUM    CARD
INPUT FORMAT    FREEFIELD
# OF CASES      585
VAR LABELS      IDENT,IDENTIFICACION DEL CLIENTE/
                NCUEN,NUMERO DE CUENTA/VAR014,ESTADO CUENTA
VALUE LABELS    IDENT (0) PARTICULAR (1) FISCAL (2) COOPERATIVA/
                VAR014 (0) PENDIENTE (1) CANCELADA (2) ATRASADA
                (3) NO INFORMAN
MISSING VALUES  VAR014 (3)
CROSSTABS       IDENT BY VAR014
```

```
STATISTICS      1,3,4
READ INPUT DATA
.
.
.
SAVE FILE
FINISH
/*
```

Ejemplo N° 2.-

```
1                16
//SPSSH PROC
//.
//.
//.
//A EXEC SPSSH
//FT03F001 DD  DSN=ARS,UNIT=2400,LABEL=(1,SL),DISP=OLD,
//            DCB=BLKSIZE=4000,VOL=SER=809
//
//SYSIN DD*
RUN.NAME        LECTURA ARCHIVO DEL SISTEMA
GET FILE        CUENTAS
CONDESCRIPTIVE  VAR015 TO VAR041
STATISTICS      ALL
FINISH
/*
```

5. MANEJO DE DATOS

5.1. Recodificación y Transformación de Variables

A menudo, durante el análisis estadístico, surge la necesidad de modificar o transformar algunas o todas las variables del archivo. Por ejemplo, puede ser de interés visualizar la distribución de alguna variable y para esto es necesario categorizar, (es decir, distribuir en clases) los valores de la variable. También puede que se quiera construir un índice que incluya el efecto conjunto de algunas variables del archivo o en el caso de una correlación, puede ser importante considerar alguna función de las variables (correlaciones de tipo no-lineal).

El sistema proporciona al usuario la posibilidad de hacer estas transformaciones en una forma rápida y sencilla.

5.1.1. Recodificación

Este proceso se controla mediante la tarjeta RECODE, cuyo formato general es:

```
1                16
RECODE   lista de variables (lista de valores=nuevo
         valor) (lista de valores=nuevo valor)...../
         lista de variables (lista de valores=nuevo
         valor).....
```

La lista de variables admite especificaciones del tipo VAR1 TO VAR2 al igual que la lista de variables de la tarjeta PRINT FORMATS.

La especificación de valores a recodificar tiene una estructura que presenta distintas opciones de acuerdo

al tipo de variable a recodificar.

Su formato general es:

$$(a_1, a_2, a_3, \ldots\ldots\ldots\ldots a_u = b)$$

Donde las a_i constituyen la lista de valores a recodificar con el valor b.

Ejemplo:

```
RECODE      NUME(1,3,5=0)(2,4,6=1)/ALFA('0','1','A'='0')
              ('2','B'='1')
```

Los valores alfanuméricos deben ir entre paréntesis.

Las variables numéricas continuas pueden ser recodificadas en categorías. En este caso se usa la palabra clave THRU para especificar los límites del intervalo.

Ejemplo:

```
RECODE      NUME(1 THRU 10=1)(10 THRU 20=2).....
            ....(100 THRU HIGHEST=11)
```

Para facilitar la recodificación de variables alfanuméricas a valores numéricos, existe una instrucción especial: CONVERT.

La especificación

```
1           16
RECODE      ALFA(CONVERT)
```

produce el mismo resultado que:

```
RECODE      ALFA('1'=1) ('2'=2) ('3'=3) ('4'=4) ('5'=5)
              ('6'=6) ('7'=7) ('8'=8) ('9'=9) ('-'=11)
              ('E'=12) (' '=13)
```

Si la instrucción RECODE es usada conjuntamente con la instrucción SAVE FILE produce un archivo del sistema en el cual las variables que han sido recodificadas, aparecen los nuevos valores.

Existe también la posibilidad de recodificar, sólo temporalmente algunas o todas las variables del archivo, para usarl s con esos nuevos valores en un proceso específico.

Para esto se coloca un asterisco precediendo la palabra RECODE. El resto de la especificación es igual. Esta recodificación sólo tiene efecto para el proceso donde se especifica.

Todos los RECODE permanentes deben preceder a los *RECODE y a las primeras tarjetas de procesos, los *RECODE preceden a las tarjetas del proceso para el cual quiere usarse esa recodificación.

Ejemplo:

```
|                16
//SPSSH PROC
//...
//...
//A EXEC SPSSH
//FT03F001 DD  DSN=USUARIO,UNIT=2400,LABEL=(1,SL),DISP=OLD,
//             DCB=(RECFM=FB=80,BLKSIZE=800),VOL=SER=325
//FT04F001 DD  DSN=SISTEMA,UNIT=2400,LABEL=(1,SL),DISP=(NEW,KEEP),
//             DCB=BLKSIZE=4000,VOL=SER=328
//SYSIN    DD *
RUN NAME       USO DE RECODE Y GENERACION ARCHIVO
FILE NAME      ENSOC ENCUESTA SOCIAL
VARIABLE LIST  EDAD,SEXO,INGRESO,EDUC,VAR001 TO VAR010
INPUT NAME     TAPE
INPUT FORMAT   FIXED(5X,F3.0,5X,A1,5X,F6.0,A1,10F4.0)
# OF CASES     2500
PRINT FORMATS  SEXO,EDUC(A)
RECODE         SEXO('H'=1) ('M'=2)/EDUC ('P'=1) ('S'=2) ('I'=3)
               ('T'=4) ('P'=5) ('N'=0) ('X'=-1)
VALUE LABELS   EDUC (1) PRIMARIA (2) SECUNDARIA (3) INDUSTRIAL (4)
               TECNICA (5) PROFESIONAL (0) ANALFABETO (-1)
               NO CONTESTO
MISSING VALUES EDUC (-1)/INGRESO(-99)
CODEBOOK       SEXO,EDUC
OPTIONS        4
```

```
STATISTICS      3,4
READ INPUT DATA
*RECODE         INGRESO (0 THRU 2500=1) (2500 THRU 10000=2)
                (10000 THRU 25000=3) (25000 THRU 50000=4)
                (50000 THRU HIGHEST=5) (-99=0)
CODEBOOK        INGRESO
OPTIONS         5
STATISTICS      ALL
CONDESCRIPTIVE INGRESO,VAR001 TO VAR010
STATISTICS      ALL
SAVE FILE
FINISH
/*
```

Se tiene aquí un proceso de lectura de archivo del usuario desde cinta (se debe usar tarjeta FT08), y un proceso de grabación de un archivo del sistema (tarjeta //FT04). La primera instrucción de recodificación permanece en efecto durante todo el programa y así en el archivo del sistema los valores de las variables SEXO y EDUC serán 1,2 y 1,2,3,4,5, 0,-1 respectivamente.

La recodificación temporal de INGRESO, se usa sólo en el subprograma CODEBOOK y se efectúa para obtener un histograma (Opción 5), los estadísticos deseados se obtienen de los valores originales mediante el subprograma CONDESCRIPTIVE que se especifica a continuación del anterior.

5.1.2. Transformación de Variables por medio de Expresiones Algebraícas

El sistema permite transformar variables existentes o crear nuevas variables por medio de expresiones algebraícas. La instrucción es:

```
 1                  16
COMPUTE  variable a calcular=expresión algebra
```

La variable a calcular puede ser entonces una variable existente o una nueva variable que se añadirá al archivo como si hubiera sido declarada a continuación de la úl-

tima variable en la tarjeta VARIABLE LIST.

La expresión algebraíca puede estar compuesta por nombres de variables existentes, operadores algebráicos, funciones matemáticas y/o constantes numéricas.

Al igual que en el caso de la instrucción RECODE, existe el *RECODE cuyos resultados son aplicables solamente al proceso estadístico que precede, no teniendo ningún efecto sobre el resto del programa.

Los operadores algebraícos son:

+	suma
-	resta
*	multiplicación
/	división
**	exponenciación

Las funciones matemáticas son:

SQRT	Raíz cuadrada
LN	Logaritmo Natural
LG10	Logaritmo base 10
EXP	Exponenciación
SIN	Seno +
COS	Coseno +
ATAN	Arcotángente +
RND	Redondeo a entero

+ argumento en radianes

Ejemplo:

```
1         10
COMPUTE   NVAR=VARA+SIN(VARB+VARC)+LN(2*VARD)+10
```

Convenciones para construir expresiones aritméticas

1. No pueden aparecer dos operadores juntos.
2. Se pueden usar paréntesis para indicar el orden de las operaciones.

3. Cuando el orden de las operaciones no ha sido completamente especificado por medio de paréntesis, las operaciones se hacen en el siguiente orden:

   a) Primero se evalúan las funciones

   b) En segundo lugar se efectúan las exponenciaciones

   c) En tercer lugar las multiplicaciones y divisiones de izquierda a derecha

   d) En último lugar las sumas y restas, también de izquierda a derecha

5.1.2.1. Tarjeta COUNT

A menudo el usuario está interesado en construir una variable cuyo valor refleje el número de respuestas de un cierto tipo, dadas en un cuestionario.

Esta variable puede construirse fácilmente mediante la instrucción COUNT. Su formato es:

```
I'          16
COUNT       variable a calcular=lista de variables
            (valores)/variable a calcular=lista de
            variables(valores)/
```

Ejemplo:

```
1           16
COUNT       ACTITUD=VAR001 TO VAR009 (1)
RECODE      ACTITUD (0 THRU 3=1) (4 THRU 6=2) (7 THRU 9=
VALUE LABELS ACTITUD (1) EN CONTRA (2) INDECISA (3) A FAV
```

El resultado en la nueva variable ACTITUD, para cada caso, es igual al número de valores 1 que tengan las variables VAR001, VAR002.....VAR009, para cada caso.

5.1.3. Transformación de Variables con Asignación Lógica

La tarjeta IF permite efectuar transformaciones de variables en la misma forma que la tarjeta COMPUTE.

La diferencia es que además permite hacer estas transformaciones sobre condiciones lógicas que son especificadas por el usuario. Su formato general es:

```
1        16
IF       (expresión lógica) variable a calcular=
         expresión aritmética
```

El cálculo se efectúa cuando el valor de la expresión lógica es Verdadero.

La expresión lógica está constituída por uno o más conjuntos de relaciones que a su vez se componen de dos expresiones aritméticas formadas como en el caso de COMPUTE que se comparan algebraícamente por medio de un operador de relación.

Estos operadores son:

GE	Mayor o igual
LE	Menor o igual
GT	Mayor que
LT	Menor que
E	Igual
NE	No igual

Ejemplo:

```
1        16
IF       ((VARA-VARC) GT   ZETA)) VENT=(A*B)-5
```

En este ejemplo cuando la diferencia entre VARA y VARC sea mayor que ZETA, la variable VENT tomará el valor de A por el valor de B menos 5.

Para combinar relaciones en una expresión lógica más compleja, existen dos operadores lógicos; AND y OR. Así la expresión lógica puede tener el siguiente formato:

(relación operador lógico relación)

Estos operadores combinan los valores lógicos de las relaciones del siguiente modo:

- AND el resultado es verdadero si y sólo si ambas relaciones son verdaderas
- OR el resultado es falso si y sólo ambas relaciones son falsas

Ejemplo:

```
1          16
IF         (VARA EQ 9 OR (VARA GT 0 AND VARA LT VARB))
           VARD=1
```

VARD toma el valor 1 cuando VARA=9 ó 0<VARA<VARB

Finalmente, existe otro operador para expresiones lógicas: NOT. Su efecto es invertir el valor de la expresión lógica que precede.

Ejemplo:

```
1          16
IF         (NOT(VARA GT 4 AND VARA LT 7))
           VARD=VARC**2
```

VARD toma el valor de VARC al cuadrado si no sucede que 4 sea menor que VARA y esta sea menor que 7

Al igual que en el caso de RECODE y COMPUTE, existe la posibilidad de hacer este tipo de transformación sólo temporalmente. Es decir, para ser utilizados en un proceso específico y sin tener efecto en la grabación de archivos del sistema. Se hace de la misma forma que en los casos anteriores, poniendo un asterisco antes de IF.

Ejemplo:

```
1          16
*IF        (A GT B AND (NOT (6 LT B))) E=AxB-C
```

5.1.4. Descripción e Inicialización de Variable Transformadas

5.1.4.1. Valores Faltantes (MISSING VALUES)

Cualquier variable a la cual se le han asignado valores por medio de las instrucciones COMPUTE, IF, COUNT y RECODE, puede tener un conjunto de valores faltantes asociados a ella. Estos valores se especifican por medio de una tarjeta MISSING VALUES que en este caso debe ponerse después de las tarjetas de modificación y antes de las tarjetas de procesos estadísticos.

Existe además una instrucción especial ASSIGN MISSING, que permite al usuario asignar un valor faltante a las variables creadas por medio de las instrucciones COMPUTE e IF. Se le asigna el valor faltante a la nueva variable, cada vez que alguna de las variables que intervienen en las instrucciones de transformación tiene un valor faltante.

Su formato es:

```
1                16
ASSING MISSING   lista de variables (valor faltante)/.....
                 lista de variables (valor faltante).....
```

5.1.4.2. Inicialización

Se puede presentar un problema cuando una variable es creada por medio de una transformación IF ó *IF. Puede darse el caso de que el valor de la variable quede indefinido. Para proporcionar resultados consistentes el sistema inicializa las variables que son creadas, asignándoles el primer valor faltante declarado para esa variable. Si el usuario no ha declarado valores faltantes para la variable, esta se inicializa asignándole el valor que tiene una constante llamada TMISS en el momento que la variable es declarada. Este valor es originalmente cero, pero el usuario puede cambiarlo mediante la instrucción.

```
1          16
THISS      valor
```

5.1.4.3. Descripción

Se pueden definir nuevos rótulos para las variables transformadas, nuevos rótulos de valores, nuevos formatos de impresión, etc. de la misma manera que se hacía al definir un archivo. Es también útil definir un nuevo nombre de archivo cuando este va a ser grabado como archivo del sistema.

Estas nuevas tarjetas deben ir inmediatamente después de las transformaciones de variables y antes de cualquier proceso estadístico. Excepto el nombre de archivo.

Ejemplo:

```
1               16
//SPSSH PROC
//...
//...
//A EXEC SPSSH
//FT03F001 DD  DSN=SISTEMA,UNIT=2400,LABEL=(1,SL),DISP=OLD,
//             DCB=BLKSIZE=4000,VOL=SER=328
//FT04F001 DD  DSN=SISTEM2,UNIT=2400,LABEL=(1,SL),DISP=(NEW,KEEP),
//             DCB=BLKSIZE=4000,VOL=SER=325
//SYSIN DD*
RUN NAME        TRANSFORMACION DE VARIABLES
GET FILE        ENSOC
FILE NAME       ENSOC2,SEGUNDA VERSION DE ENSOC
COMPUTE         IND=INGRESO-(VAR002*2000)
IF              (IND LE 2000) INREAL=1
IF              ((IND GT 2000) AND (IND LE 8000)) INREAL=2
IF              ((IND GT 8000) INREAL=3
VAR LABELS      INREAL,INGRESO REAL AJUSTADO POR NUCLEO FAM
VALUE LABELS    INREAL (1) BAJO VITAL (2) 2 A 4 VITALES
                (3) + DE 4 VITALES
PRINT FORMATS   IND,INREAL (0)
CODEBOOK        IND,INREAL
OPTION          5
SAVE FILE
FINISH
/*
```

El nuevo archivo del sistema tiene como nombre LNSOC2, incluye las variables IND e INREAL con toda la información sobre rótulos, valores faltantes, etc. que se incluyen en el programa.

5.1.5. Limitaciones en la Recodificación y Transformación de Variables

5.1.5.1. Limitaciones de la Codificación

1. No más de 250 elementos pueden aparecer en una especificación (*)RECODE. Cuentan como elementos:

   - los nombres de variables
   - palabras claves (TO,THRU,LOWEST,HIGHEST)
   - valores individuales
   - paréntesis izquierdo y derecho y
   - signo igual

2. El número de valores individuales que aparecen a la izquierda del signo igual (excluyendo aquellos del tipo X THRU Y), más el número de listas de variables no puede exceder de 200.

3. El número de veces que aparece una especificación del tipo X THRU Y, más el número de veces que aparece la especificación (CONVERT), más el número de listas de variables, no puede exceder de 200.

4. No pueden ser recodificados más de 140 variables.

Estas cuatro restricciones son independientes y pueden estar todas en su máximo valor sin producir error.

+ (*)RECODE significa RECODE y/o *RECODE.

5.1.5.2. Limitaciones Mixtas

1. La suma de todos los operadores en las tarjetas (*)COMPUTE, (*)IF y (*)SELECT IF (ver punto 5.2.2.) que preceden a una tarjeta de procedimiento no pueden exceder de 1000.

   Cada operador de relación (EQ,NE,GT,LT, LE y GE) cuenta como uno. Cada operador lógico (AND,OR o N cuenta como uno. Cada operador aritmético (**,*,+,-,) cuenta como uno.

2. La suma del número de (*)RECODE más el número de (*)COMPUTE más dos veces el número de (*)IF o (*)SELECT IF no puede exceder de 250.

5.2. Selección de Casos de un Archivo

El sistema proporciona distintas facilidades para seleccionar casos dentro de un archivo.

5.2.1. Muestra al Azar

Para el caso en que el usuario tenga u archivo muy grande y desee tener una estimación de los parámet estadísticos, el sistema proporciona la posibilidad de generar una muestra al azar del archivo. Esto se hace usando la instrucción:

```
1          16
SAMPLE     factor
```

Factor es un número decimal menor que y que indica el porcentaje de los cas a incluirse en la muestra.

5.2.2. Selección sobre Condiciones Lógicas

Si el usuario quiere procesar sólo aq llos casos del archivo para los cuales se cumple cierta condi puede hacerlo mediante la instrucción:

```
1          16
SELECT IF  (expresión lógica)
```

La expresión lógica tiene las mismas reglas de formación y los mismos operadores que en el caso de la tarjeta IF.

5.2.3. Selección de Casos
Generación de Archivos

Cuando se usan las instrucciones SAMPLE y/o SELECT IF conjuntamente con la instrucción SAVE FILE el resultado es que el archivo generado contiene solo el porcentaje de casos especificados en la tarjeta SAMPLE y/o sólo los que cumplen la condición especificada en SELECT IF.

Puede que no sea este el deseo del usuario y por este motivo existe el *SAMPLE y el *SELECT IF que permiten la selección para un proceso específico y sin que tengan efectos para la grabación.

La selección permanente (sin asterisco) debe hacerse antes del primer proceso estadístico. La selección temporal debe hacerse inmediatamente antes del proceso en el cual se quiere usar.

5.3. Ponderación de los Casos

El sistema proporciona la posibilidad de ponderar los casos. Este procedimiento es aplicable al caso de muestras mal dimensionadas o de muestreo estratificado.

Esta ponderación puede realizarse de dos maneras. Incluyendo en cada caso una variable de ponderación, o generando una nueva variable, por medio de transformaciones, a la cual se le asigne esta función.

Ejemplo:

La variable PON de cada caso es la función de peso

```
//SPSSH PROC
:
RUN NAME        PRUEBA
VARIABLE LIST VAR001 TO VAR010,VARB PON,VAR2
:
:
WIEGHT          PON
```

Un método de generar la función de ponderación podría ser:

```
//SPSSH  PROC
:
:
RUN NAME        PRUEBA
VARIABLE LIST   VARA,VAR001 TO VAR010,VARB,VARZ  _
:
:
IF              (VARA LT 3) PON=2
IF              (VARA GE 3 AND VARA LT 7) PON=1.5
IF              (VARA GE 7) PON=1
WEIGHT          PON
```

Existen dos versiones del WEIGHT. Una permanente y otra temporal. (*WEIGHT).

5.4. Orden de Precedencia en Tarjetas de Manejo de Datos

El orden es:

Tarjeta	Rango	
SAMPLE	1	(mayor rango)
SELECT IF	2	
RECODE	2	
COMPUTE	2	
IF	2	
WEIGHT	3	
*SAMPLE	4	
*SELECT IF	5	
*RECODE	5.	
*COMPUTE	5	
*IF	5	
*WEIGHT	6	(menor rango)

Dentro de un mismo rango las tarjetas son intercambiables.

6. TARJETAS DE CONTROL DEL SISTEMA

6.1. KEYPUNCH

Esta tarjeta se usa solo cuando las tarjetas del sistema han sido perforadas en una máquina IBM-026. Debe preceder a la tarjeta RUN NAME. Su formato es:

```
1          16
KEYPUNCH   026
```

6.2. PRINT BACK

Normalmente las tarjetas de control del sistema aparecen en los listados. Si se quiere suprimir la impresión de estas en los listados, se usa la tarjeta PRINT BACK. Su formato es:

```
1          16
PRINT BACK NO
```

Precede a la tarjeta RUN NAME y sigue a la tarjeta KEYPUNCH.

6.3. NUMBERED

Esta tarjeta permite reservar los últimos 8 campos de cada tarjeta de control del sistema, para perforar códigos de identificación. Es especialmente útil para programas grandes. Si se usa debe ser la primera tarjeta de control del sistema. Su formato es:

```
1          16
NUMBERED   YES
```

```
1               16                                                  72
//SPSSH PROC
//
//
//SYSIN DD *
NUMBERED        YES                                          ..... 001
RUN NAME        PRUEBA                                       ..... 002
VARIABLE LIST   VARA,VAR001 TO VAR025,VARB,VARC,.....        ..... 003
                VARD,VAR026 TO VAR050                        ..... 003.1
INPUT MEDIUM    CARD                                         ..... 004
                                                                 :
                                                                 :
READ INPUT DATA                                              ..... nn6
                 Datos
                                                                 :
                                                                 :
FINISH                                                       ..... nn7
/*
```

6.4. COMMENT

Permite intercalar comentarios en cualquier lugar entre las tarjetas de control del sistema. Esta tarjeta puede ser continuada en tarjetas subsecuentes (tantas como sea necesario), dejando el campo de control en blanco. Su formato es:

```
1       16
COMMENT 'cualquier comentario que el usuario
        desee intercalar'
```

Se puede intercalar cualquier cantidad de comentarios.

6.5. DOCUMENT

Esta tarjeta es análoga a la tarjeta COMMENT pero su función es permitir documentar más bien el archivo que el programa. Por ejemplo, si se hace alguna transformación de variables o alguna recodificación, puede incluirse, en el archivo, una explicación de los criterios usados.

La información que se incluye en los archivos a través de la tarjeta DOCUMENT puede ser listada en procesos posteriores. Su formato es:

```
1          16
DOCUMENT   'cualquier documentación del archivo
           que el usuario desee agregar a éste'
```

Al igual que en el caso de COMMENT, puede ser continuada la documentación en tarjetas sucesivas (tantas como se necesario) y no hay limitación en el número de DOCUMENT que se especifiquen.

6.6. DUMP

Esta tarjeta permite examinar la información descriptiva de un archivo generado por el sistema SPSS. Su formato es:

```
1     16
DUMP  'lista de claves'
```

Esta lista de claves puede contener una o más de las siguientes claves:

- VARLIST : causa que se imprima una lista de todas las variables del archivo, en el orden en que fueron grabadas e incluyendo aquellas que se han creado por transformaciones
- VARINFO : causa que se imprima una lista de los formatos de impresión y valores faltantes para cada variable del archivo
- SUBDIRECTORY : causa que se imprima una lista de todos los subarchivos y el número de casos en cada uno. (Ver punto 7.)
- LABELS : causa que se imprima una lista de todos los rótulos asociados al archivo y a sus variables

- DOCUMENT : causa que se imprima toda la documentación que ha sido incluida en el archivo a través de las tarjetas DOCUMENT

- SORTVARS : causa que se imprima una lista de todas las variables ordenadas por orden alfabético

- LABCARDS : causa que se impriman los rótulos de variables y los rótulos de valores de variables tal como se especificaron originalmente

- COMPLETE : causa que se imprima para cada variable, el nombre, rótulos asociados, valores faltantes y formatos de impresión.

7. SUBARCHIVOS

Los casos en un archivo pueden ser divididos formalmente en hasta 100 subarcivos. Esto permite efectuar procesos estadísticos sobre grupos de datos en forma independiente. Estos grupos de datos, desde el punto de vista estadístico, pueden considerarse como subpoblaciones.

7.1. Especificación de Subarchivos

Los subarchivos se especifican mediante la tarjeta SUBFILE LIST, que tiene una estructura similar a la tarjeta VARIABLE LIST. Los nombres de los subarchivos pueden ser de hasta 8 caracteres alfanuméricos, el primero alfabético. La diferencia que presentan con respecto a los nombres de variables es que deben ser únicos en término de los cuatro primeros caracteres con respecto a los nombres de los otros subarchivos. También deben ser únicos con respecto a los nombres de variables.

La tarjeta SUBFILE LIST va inmediatamente después de la tarjeta VARIABLE LIST.

Su formato es:

```
                1          16
SUBFILE LIST    nombre subarchivo1, nombre subarchivo2,
                ..........,nombre subarchivo n,.....
```

Los subarchivos deben nombrarse uno a uno.

7.2. Número de casos en los Subarchivos

Cuando el usuario ha especificado estructura de subarchivos en su archivo, la tarjeta # OF CASES debe indicar el número exacto de casos en cada subarchivo.

Ejemplo:

```
1                16
SUBFILE LIST     SUB1,SUB2,SUB3,SUB4,SUB5
.............      |    |    |    |    |
# OF CASES      1523,1407,1617,1726, 593
```

El archivo contiene 6866 casos que están divididos (mediante algún criterio del usuario) en cinco grupos que constituyen otros tantos subarchivos. Estos son:

- Del caso 1 al 1523 contituye el subarchivo SUB1.
- Los 1407 casos siguientes constituyen el subarchivo SUB2.
- Los siguientes 1617 el subarchivo SUB3.
- Los siguientes 1726 el subarchivo SUB4.
- Los últimos 593 el subarchivo SUB5.

7.3. Procesamiento de Subarchivos

Cuando el archivo está dividido en subarchivos, el usuario debe informar al sistema como se van a procesar estos subarchivos. Esto se hace a través de la tarjeta PROCESS SBFILES.

Las opciones disponibles son:

- Se procesa cada archivo por separado.
- Un subarchivo o un conjunto particular de subarchivos pueden ser seleccionados de manera que los procesos se ejecuten sólo para esos subarchivos.
- Los subarchivos pueden ser agrupados en conjuntos más grandes de casos para efectos de un proceso específico.
- La estructura de subarchivos puede ser totalmente ignorada.

Las especificaciones de estas opciones se explican más claramente a continuación. Tomemos el archivo del ejemplo anterior.

- Si el usuario quiere procesar cada subarchivo en forma independiente, tendría que especificar

```
1              16
PROCESS SBFILESEACH
```

- Si se quiere procesar sólo los subarchivos SUB1 y SUB2 en forma independiente

```
1                16
PROCESS SBFILES(SUB1)(SUB2)
```

- Si se quiere combinar SUB3 y SUB4 en un grupo SUB2 y SUB5 en otro y procesar aparte SUB1

```
1                16
PROCESS SBFILES(SUB1),(SUB3,SUB4),(SUB2,SUB5)
```

- Finalmente, si se quiere ignorar la estructura de subarchivos, se especifica

```
1                16
PROCESS SBFILESALL
```

La tarjeta PROCESS SBFILES se coloca antes del primer set de tarjetas de procesos y permanece en efecto hasta que encuentra otra declaración PROCESS SBFILES.

7.4. Ubicación y Precedencia de Tarjetas de Control

A continuación se da una lista del orden en que deben aparecer las tarjetas de control del sistema SPSS vistas hasta este momento.

Precedencia	Tarjeta	Status
0	KEYPUNCH	Obligatoria si se usa IBM-026
2	PRINT BACK	Opcional
3	RUN NAME	Opcional
4	GET FILE	Obligatoria si se lee archivo del sistema
5	FILE NAME	Obligatoria si se graba archivo del sistema
6	VARIABLE LIST	Obligatoria si se lee archivo del usuario
7	SUBFILE LIST	Obligatoria si se lee archivo usuario con subarchivos

8	INPUT MEDIUM	Obligatoria si se lee archivo del usuario
9	INPUT FORMAT	
10	# OF CASES	
11	PROCESS SBFILES	Obligatoria si el archivo tiene estructura de subarchivo
12	Tarjetas de procesc	

8. PROCEDIMIENTOS ESTADISTICOS II
Relaciones entre dos o más Variables

Luego de haber examinado la distribución de cada una de las variables del archivo, el siguiente paso, normalmente, es tratar de establecer las relaciones que pueden existir entre estas variables.

Dados los tipos de variables y el propósito del análisis, uno o más procedimientos pueden ser seleccionados. El sistema SPSS dispone de tres procedimientos para este fin, dos de ellos (CROSSTABS y FASTABS) para confeccionar tablas de frecuencia de doble entrada o más (mediante el uso de variables de control), y el tercero (SCATTERGRAM) que grafica variables continuas.

Los dos subprogramas de tabulaciones realizan básicamente las mismas funciones y ambos vienen complementados con test de significancia y medidas de asociación que pueden ser seleccionados por el usuario. Una diferencia es que FASTABS está restringido a variables numéricas discretas, mientras que CROSSTABS puede operar con cualquier tipo de variable (alfanumérica discreta). Esta limitación de FASTABS hace que tenga mayor rapidez de proceso.

Naturalmente la restricción de variables discretas para CROSSTABS y numéricas discretas para FASTABS puede obviarse facilmente a través de la instrucción RECODE.

8.1. Subprograma CROSSTABS

Este subprograma se usa normalmente para hacer tabulaciones de variables alfanuméricas. Para variables numéricas discretas (o continuas en clases), es más eficiente el subprograma FASTABS.

CROSSTABS no requiere que el usuario especifique el número de categorías de las variables a tabular.

8.1.1. Tarjeta de Procedimiento

La tarjeta de especificación lleva en su campo de control la clave CROSSTABS. En el campo de especificación se indican las tablas requeridas indicando las dos variables a tabular separándolas por la palabra BY.

Ejemplo:

```
CROSSTABS     VALOR BY COSTO
```

Se pueden pedir varias tablas en una misma tarjeta, separando las especificaciones con un slash.

Ejemplo:

```
CROSSTABS     VALOR BY COSTO/HMBRE BY HMAQ/.....
```

Además cuando se quiere tabular un conjunto de variables con una o más variables, se puede especificar una lista de tablas.

Ejemplo:

```
CROSSTABS     VAR1,VAR2,VAR4 BY VAR5,VAR6/VAR7 TO
              VAR1Ø,VAR12 BY VAR25,ZETA
```

La primera especificación (o lista de tablas) produce tablas de VAR1 con VAR5, VAR1 con VAR6, VAR2 con VAR5,....., hasta VAR4 con VARó. La segunda especificación, de la misma manera, produce 10 tablas.

Ejemplos:

```
1                16
//SPSSH PROC
//.........
//.........
//SYSIN DD *
RUN NAME        TABULACIONES CENSO,PRIMERA MUESTRA
FILE NAME       CENSO1
VARIABLE LIST   EDAD,SEXO,EDUC,INGRESO,ESTCIV,NHIJOS
```

```
1                16
SUBFILE LIST     STGO,VALPO,CONCE
INPUT MEDIUM     CARD
INPUT FORMAT     FREEFIELD
# OF CASE        5000
RECODE           INGRESO (0 THRU 2100=1)(2100 THRU 10500=2)(10500
                 THRU 21000=3)(21000 THRU 42000=4)(42000 THRU
                 HIGHEST=5)(-99=0)/EDAD (0 THRU 17=1)(18 THRU
                 25=2)(25 THRU 35=3)(35 THRU 45=4)
                 (45 THRU HIGHEST=5)
PROCESS SBFILESEACH
COMMENT          LOS SUBARCHIVOS SE PROCESAN EN FORMA INDEPENDIENTE
CROSSTABS        EDAD TO EDUC BY INGRESO
COMMENT          SE PIDEN TABLAS DE DOBLE ENTRADA DE EDAD-INGRESO,
                 SEXO-INGRESO,EDUC-INGRESO
READ INPUT DATA  }
.............    }
.............    }  Datos en tarjetas
.............    }
CROSSTABS        EDAD BY SEXO/INGRESO BY EDUC BY EDAD
COMMENT          SE PIDE TABLA EDAD-SEXO/SE PIDE TABLA A TRES
                 DIMENSIONES:RESULTA UNA TABLA INGRESO-EDUC POR
                 CADA VALOR DE LA VARIABLE DE CONTROL EDAD
PROCESS SBFILESALL
COMMENT          AHORA SE IGNORA LA ESTRUCTURA DE SUBARCHIVO
CROSSTABS        INGRESO BY EDUC BY EDAD BY SEXO
COMMENT          SE PIDE UNA TABLA A CUATRO DIMENSIONES,RESULTADO:
                 UNA TABLA INGRESO-EDUC POR CADA UNA DE LAS COMBI
                 NACIONES DE VALORES DE EDAD-SEXO,VARIANDO PRIMERO
                 EDAD.(TODOS LOS VALORES DE EDAD PARA CADA VALOR
                 DE SEXO)
FINISH
/*
```

EJEMPLOS
FILE DATOS (CREATION DATE = 03/13/73) FICTICIOS DE EJEMPLO

C R O S S T A B U L A T I O N O F SEXO BY INGRESO

SEXO \ INGRESO	1.00	2.00	3.00	ROW TOTAL
0.0	13	18	5	36
	36.1	50.0	13.9	60.0
	48.1	69.2	71.4	
	21.7	30.0	8.3	
1.00	14	8	2	24
	58.3	33.3	8.3	40.0
	51.9	30.8	28.6	
	23.3	13.3	3.3	
COLUMN TOTAL	27	26	7	60
	45.0	43.3	11.7	100.0

CHI SQUARE = 2.80427 WITH 2 DEGREES OF FREEDOM
CRAMER'S V = 0.21933
CONTINGENCY COEFFICIENT = 0.21416
KENDALL'S TAU B = -0.20356
KENDALL'S TAU C = -0.21778
GAMMA = -0.37121
SOMER'S D = -0.18267

EJEMPLOS

FILE DATOS (CREATION DATE = 03/13/73) FICTICIOS DE EJEMPLO

* * * * * * * * * * * * * * * * * * * CROSSTABULATION OF PROF BY INGRESO * * * * * * * * * * * * * * * * * * * PAGE 1 OF

PROF	INGRESO			ROW TOTAL
COUNT / ROW PCT / COL PCT / TOT PCT	MENOS DE 5 VIT (1.00)	ENTRE 5 Y 15 (2.00)	MAS DE 15 (3.00)	
EMP.U OBRERO (1.00)	25 / 59.5 / 92.6 / 41.7	16 / 38.1 / 61.5 / 26.7	1 / 2.4 / 14.3 / 1.7	42 / 70.0
TECNICO (2.00)	2 / 20.0 / 7.4 / 3.3	6 / 60.0 / 23.1 / 10.0	2 / 20.0 / 28.6 / 3.3	10 / 16.7
PROFESIONAL UNI (3.00)	0 / 0.0 / 0.0 / 0.0	4 / 50.0 / 15.4 / 6.7	4 / 50.0 / 57.1 / 6.7	8 / 13.3
COLUMN TOTAL	27 / 45.0	26 / 43.3	7 / 11.7	60 / 100.0

CHI SQUARE = 21.72218 WITH 4 DEGREES OF FREEDOM
CRAMER'S V = 0.42546
CONTINGENCY COEFFICIENT = 0.51556
KENDALL'S TAU B = 0.51103
KENDALL'S TAU C = 0.40333
GAMMA = 0.82313
SOMER'S D = 0.45107

El formato general de la tarjeta CROSSTABS es:

```
1                16
CROSSTABS        {lista de variables}BY {lista de variable
                 BY.....BY {lista de variables}/{lista de
                 variables}BY...../.....
```

Las listas de variables admiten especificaciones del tipo VARA TO VARZ.

8.1.2. Opciones

El listado normal de CROSSTABS, entrega en cada celda de la tabla, la frecuencia absoluta, porcentaje de la columna, de la fila y del total. Estos porcentajes pueden suprimirse, si el usuario lo desea, a través de opciones disponibles en el subprograma (opciones 3,4 y 5).

Las opciones disponibles son:

Opción 1. Causa que se incluyan en las tabulaciones los valores declarados como valores faltantes.

Opción 2. Causa que no se imprima ningún tipo de rótulo.

Opción 3. Causa que no se impriman los porcentajes por fila.

Opción 4. Causa que no se imprimar los porcentajes por columnas.

Opción 5. Causa que no se impriman los porcentajes del total.

Si no se especifican opciones el subprograma asume:

a. Se excluyen de las tablas los valores declarados en las tarjetas MISSING VALUES.

b. Se imprimen los rótulos que se hayan especificado.

c. Se imprimen todos los porcentajes.

8.1.3. Estadísticos

Los estadísticos disponibles son:

1. Chi cuadrado *
2. Phi para tablas de 2x2, V de Cramer para tablas más grandes
3. Coeficiente de Contingencia
6. Tau B de Kendall
7. Tau C de Kendall
8. Gamma
9. D de Somer (simétrico)

* Para tablas de 2x2, se aplica el test exacto de Fisher cuando hay menos de 21 casos. Chi-cuadrado con la corección de Yates es aplicado en todas las otras tablas de 2x2. (Ver Anexo A).

8.1.4. Limitaciones del subprograma

1. No se puede hacer referencia a más de 200 variables en una tarjeta CROSSTABS. (Las referencias del tipo VARA TO VARZ cuentan sólo como 1).
2. El número máximo de valores individuales que puede tomar una variable es 250.
3. La siguiente fórmula da el número máximo de celdas que pueden ser generadas en una corrida del subprograma.

$$MAXCEL = \frac{(ESPACIO/4)}{D+2}$$

donde D es la dimensión de la tabla de mayores dimensiones declarada en el programa. ESPACIO es el mismo explicado anteriormente.

4. Un máximo de 20 listas individuales de tablas pueden ser procesadas por una tarjeta CROSSTABS. (Una lista individual de tablas se define como todas las especificaciones que terminan por '/', incluyendo la última).

5. El número máximo de entradas para una tabulación cruzada es diez. Esto significa que las variables de control pueden ser sólo ocho.

```
CROSSTABS   EDAD BY RAZA BY VARA BY VARB.....BY VARH
                           1      2          8
                                variables de control
```

8.2. Subprograma FASTABS

El subprograma FASTABS realiza básicamente las mismas funciones que CROSSTABS y los listados de ambos subprogramas son idénticos excepto por algunos estadísticos adicionales que opcionalmente calcula FASTABS; sin embargo FASTABS es mucho más rápido y puede manejar un número más grande de tablas que CROSSTABS. Esto se debe a que sólo opera con variables numéricas discretas y requiere que el usuario especifique los rangos de variación de las variables (número de categorías).

8.2.1. Tarjeta de Procedimiento

El campo de especificación de la tarjeta FASTABS consta de dos partes; la primera parte se usa para definir las variables que van a ser usadas en las tablas y sus respectivos rangos. La segunda parte define las tablas que se van a generar.

```
1          16
FASTABS    VARIABLES= lista de variables (mínimo,máximo)/
           lista de variables (mínimo,máximo)/...../
           TABLES= lista de variables  BY  lista de
           variables  BY ...../.....
```

Las listas de variables de la primera parte de la tarjeta admiten especificaciones en el tipo VARA TO VARZ.

Ejemplo:

```
FASTABAS   VARIABLES=VARA,VARC TO VARF,VARH (0,10)/
           EDUC(0,7)/...../
```

Todas las variables incluídas en la lista a la izquierda de un paréntesis deben variar entre los rangos especificados. Así VARA,VARC TO VARH sólo pueden variar entre 0 y 10. Los valores fuera de rango se excluyen de las tabulaciones, indicando el número de valores excluídos.

La segunda parte de la tarjeta es similar a la especificación de la tarjeta CROSSTABS. Las listas de tablas tienen el mismo significado que en CROSSTABS. La diferencia está en que en las listas de variables correspondientes a esta parte, las declaraciones del tipo VAR1 TO VAR5 se refrieren a variables adyacentes declaradas en las listas de la primera parte de la tarjeta. No se puede hacer referencia a variables adyacentes sólo en la tarjeta VARIABLE LIST, es necesario (y suficiente) que sean adyacentes en la primera parte de la tarjeta FASTABS.

Ejemplo:

```
1                16
FASTABS  VARIABLES=SEXO(0,1)/EDUC(0,7)/INGRESO(0,10),
         ITEM1 TO ITEM7,ITEM9,UBIC(0,5)/ITEM8(0,15)
         TABLES=SEXO,EDUC,INGRESO BY ITEM8/
         ITEM1 TO UBIC BY SEXO BY EDUC
```

Las tablas especificadas son: En la primera lista; SEXO-ITEM8,EDUC-ITEM8,INGRESO-ITEM8.

En la segunda lista se piden 9 tablas a tres dimensiones; ITEM1-SEXO/EDUC,.....UBIC-SEXO/EDUC. Esto hace un total de 72 tablas.

8.2.2. Opciones

Las opciones disponibles en FASTABS, son las mismas de CROSSTABS excepto por una opción adicional de FASTABS que es:

Opción 6 causa que no se impriman los rótulos de valores de las variables pero si los rótulos de variable

8.2.3. Estadísticos

El subprograma FASTABS tiene todos los estadísticos que posee CROSSTABS y además tiene los siguientes:

4. Lamda simétrico y asimétrico.
5. Coeficiente de incertidumbre simétrico y asimétrico.
10. Eta (η)

8.2.4. Limitaciones

1. No se puede hacer referencia a más de 100 variables en la primera parte de la tarjeta FASTABS.
2. Ninguna variable puede tener más de 250 valores individuales o categorías.
3. Un máximo de 20 listas de tablas pueden especificarse en la segunda parte de la tarjeta. Por lista de tablas se entienden las especificaciones que terminan por slach (/, incluyendo la última.
4. El número máximo de celdas que pueden ser generadas por el subprograma está dado por:

   MAXCEL= (ESPACIO/4)-(10xNVARS)

   donde NVARS es el número de variables nombradas explícita o implícitamente en la primera parte de la tarjeta y ESPACIO tiene el significado dado anteriormente. (PARM='ESPACIO').
5. El número máximo de dimensiones para una tabulación es 8. Es decir, se pueden especificar hasta 6 variables de control en cada tabla.
6. El número máximo de variables que se pueden mencionar en la segunda parte de la tarjeta FASTABS es 100. Las especificaciones del tipo VARA TO VARX cuentan solo como 2.

Nota: El número de celdas de una tabla es igual al producto del número de categorías especificadas para esas variables. Si una variable está codificada en 0, 1, 2 y 9 su rango

de variación es (0,9) y el programa considera 10 categorías (0,1,2,3,4,5,6,7,8,9). Por este, para lograr una mejor utilización de la memoria, es muy conveniente recodificar en categorías adyacentes (0,1,2,9 como 0,1,2,3).

8.3. Subprograma SCATTERGRAM

Este subprograma proporciona un gráfico, que permite visualizar el tipo de relación existente entre dos variables. El listado de este subprograma es una 'nube de puntos' encuadrada por dos ejes, en los cuales se representan las dos variables.

El gráfico consiste de 51 posiciones verticales y 101 posiciones horizontales. El subprograma ajusta, automáticamente, una escala a cada eje, de acuerdo al rango de la variable correspondiente. Además permite que el usuario fije una o ambas escalas, especificando un rango dado para una o ambas variables. En este caso los valores que caen fuera del rango especificado son excluidos del gráfico y son listados bajo este con el mensaje EXCLUDED VALUES.

El subprograma además incluye algunos estadísticos básicos y la posibilidad de ajustar una recta a la 'nube de puntos' por el método de mínimos cuadrados.

8.3.1. Tarjeta de Procedimiento

El campo de especificación de la tarjeta de procedimiento puede tener tres tipos de formatos diferentes.

a)

```
1                16
SCATTERGRAM  lista de variables
```

Esta lista de variables admite todas las convenciones ya mencionadas. Esta especificación produce gráficos entre la primera variable de la lista y cada una de las siguientes, luego entre la segunda variable de la lista y cada una de las siguientes y así hasta la penúltima variable de la lista con la última.

Ejemplo:

```
1                16
SCATTERGRAM  VAR001 TO VAR005,VAR007,ZETA
```

En este caso se generarían 21 gráficos. El número de gráficos se determina por $NG=(n^2-n)/2$, donde n es el número de variables en la lista. El primer gráfico es con VAR001 como variable vertical y VAR002 como variable horizontal, el segundo con VAR001 vertical y VAR003 horizontal y el último con VAR007 vertical y ZETA horizontal.

b)

```
1                16
SCATTERGRAM  lista de variables WITH lista de
             variables/.........
```

En el segundo tipo de formato, se usa la palabra clave WITH para producir gráficos sólo entre las variables mencionadas antes y después de WITH. El slach '/' es usado para separar listas de gráficos.

Ejemplo:

```
1                16
SCATTERGRAM  VAR001,VAR002 WITH VAR007 TO VAR009/....
```

La primera lista de gráficos (termina en un /) generaría 6 gráficos, 3 para cada variable mencionada antes de WITH. El número de gráficos está dado por NG=nxm, donde m y n corresponden al número de variables antes y después de WITH.

c)

El tercer formato es una ampliación de dos anteriores, que permite fijar una o ambas escalas del gráfico, especificando el rango de una o ambas variables. El rango se especifica poniendo el nombre de la variable y luego el mínimo y máximo de la misma, entre paréntesis. Esta especificación concierne sólo a la variable que precede a la especificación.

Ejemplo:

```
1                16
SCATTERGRAM      VAR003(0,10) WITH VARA(7,HIGHEST),
                 VAR007 TO VAR009(LOWEST,99)
```

En este caso la especificación genera 5 gráficos, todos con VAR003 vertical. La escala para VAR003 es de 0 a 10, para VARA es de 7 hasta el máximo valor leído y para VAR009 desde el menor valor leído hasta 99. Para VAR007 y VAR008 no se ha especificado rango, por lo tanto la escala se ajusta en forma automática para cada una de estas variables.

El formato general sería:

```
1                16
SCATTERGRAM      lista de variables (rango)/lista
                 de variables (rango) WITH lista de
                 variables (rango)/.....
```

Donde lista de variables (rango), significa una o más variables de las cuales en algunas, todas o ninguna se ha especificado un rango.

El usuario que está familiarizado con sus datos puede a menudo mejorar la calidad de los gráficos, especificando rangos tales que sean múltiplos de 10 para las variables en el eje vertical y múltiplos de 20 para las variables en el eje horizontal. De esta manera, en el listado aparecen impresos números enteros en las escalas de los ejes.

Si el usuario no conoce los datos, pero quiere que las coordenadas de las escalas sean números enteros, sin que para esto deba excluir algún caso, puede usar la opción 7 que se describe en el punto siguiente.

8.3.2. Opciones

La lista de opciones es:

Opción 1

Todos los MISSING son incluídos en el gráfico y en los estadísticos especificados en la tarjeta de estadísticos.

Opción 2.

Los MISSING VALUES son excluídos de los gráficos y de los estadísticos, lo cual significa que un caso es excluído si cualquiera de las dos variables tiene un MISSING VALUE.

Opción 3.

Suprime la impresión de los rótulos de variables.

Opción 4.

Se suprime la impresión de la grilla. Esta consiste en líneas paralelas a los ejes que dividen el gráfico en 9 rectángulos, para la mejor localización de los puntos.

Opción 5.

Imprime una grilla diagonal.

Opción 6.

Se hace un test de significancia de doble cola. (Ver Anexo Estadístico).

Opción 7.

Construye escalas en forma automática de manera tal que las coordenadas resulten números enteros.

Nota:

Si el usuario especifica un rango para una variable en particular y también selecciona la opción 7, la escala se ajusta según la primera especificación.

Si no se especifican opciones, el subprograma asume:

a) Se excluyen los casos en los cuales cualquiera de las dos variables tiene un MISSING VALUES.

b) Se imprimen los rótulos de variables.

c) Se imprime una grilla (paralela).

d) No se hace ajuste automático de escalas para producir coordenadas con valores enteros.

8.3.3. Estadísticos

Los estadísticos disponibles son:

1 Coeficiente de correlación lineal R de Pearson
2 R cuadrado
3 Significancia de R
4 Error standard de la estimación

Los otros dos estadísticos se refieren al ajuste de una recta tipo y=mx+n con X la variable horizontal e Y la vertical.

5 n, punto de corte en el eje Y
6 m, pendiente de la recta

(Ver Anexo A.Correlación).

8.3.4. Limitaciones

1. No se pueden especificar explícita o implícitamente más de 100 variables.
2. No se pueden especificar más de 20 listas de variables.
3. El número máximo de casos que el programa puede procesar está dado por:

$$MXCAS = \{(ESPACIO/4) - 2*NV\}/(1+NV)$$

NV : número de variables en la tarjeta de especificación

MXCAS : número máximo de casos, (esto es calculado por el programa ya que el Espacio está dado, si el número de casos leídos es mayor que el calculado se imprime un mensaje de error).

9. MODIFICACION DE ARCHIVOS DEL SISTEMA

Una vez que el usuario tiene sus datos grabados en un archivo del sistema, puede, a partir de ese archivo, generar un nuevo archivo del sistema modificado. Así las instrucciones de modificación, usadas conjuntamente con la instrucción SAVE FILE, permiten obtener nuevos archivos del sistema. Esto puede hacerse agregando variables o nuevos casos al archivo original del sistema, o eliminando variables o subarchivos del archivo original.

9.1. Eliminación de Variables

El usuario puede eliminar una o más variables de los casos en el archivo del sistema, generando un nuevo archivo sólo con aquellas variables en las que está interesado.

La eliminación se hace mediante la instrucción DELETE VARS o KEEP VARS, en conjunto con la instrucción SAVE FILE.

Los formatos de estas instrucciones son:

```
1                16
DELETE VARS    {lista de variables}
KEEP VARS      {lista de variables}
```

Las listas de variables admiten especificaciones del tipo VARA TO VARZ.

Las dos instrucciones KEEP VARS y DELETE VARS tienen la misma finalidad. El usar una u otra depende sólo del número de variables que se quiere retener o eliminar y de la facilidad de especificar cada uno de los dos conjuntos de variables. Si es más fácil especificar el conjunto de variables que se quiere retener, se usará la instrucción KEEP VARS. Si por el contrario es más fácil especificar el otro conjunto, se usará la instrucción DELETE VARS.

Las tarjetas DELETE VARS o KEEP VARS (cualquiera que se esté usando), se coloca inmediatamente antes de la tarjeta SAVE FILE. (Esta última es requerida en este proceso junto con las tarjetas de control del O.S. respectivas FT04 y FT03).

Es conveniente insertar un nuevo FILE NAME (siguiendo la tarjeta GET FILE), dando otro nombre al nuevo archivo.

9.2. Agregación de Variables

El usuario puede agregar variables a los casos del archivo del sistema, mediante la instrucción ADD VARIABLES, siempre que se cumplan las siguientes condiciones:

i) Debe haber una correspondencia total entre el número y orden de los casos en el archivo y el número y orden de los casos en que vienen las nuevas variables.

ii) Si el archivo del sistema tiene estructura de subarchivo, el número, orden y tamaño de cada subarchivo, así como el número de casos dentro de ellos deben estar en total concordancia en los dos archivos, el del sistema y el formado por los nuevos datos (archivo del usuario).

iii) La suma del número de variables que serán agregadas más el número de variables ya existentes en el archivo del sistema no debe exceder de 500.

Si se cumplen estas condiciones el usuario puede agregar variables a su archivo mediante la instrucción ADD VARIABLES cuyo formato es:

```
1                16
ADD VARIABLES lista de variables
```

Las convenciones para declarar variables son las mismas que las usadas en la tarjeta VARIABLE LIST. Requiriendo además que los nombres de las nuevas variables sean únicas en términos de las variables ya existentes en el archivo.

La tarjeta ADD VARIABLES va directamente a continuación de la tarjeta GET FILE. Es la segunda instrucción en ejecutarse, lo que permite programar los procesos estadísticos, que van a continuación, en base al nuevo archivo (que incluye todas las variables).

Además de la tarjeta ADD VARIABLES, deben prepararse las siguientes tarjetas:

- Tarjeta INPUT MEDIUM, especificando el dispositivo desde el cual se leerán las nuevas variables. Si los casos del archivo del usuario no están en tarjetas, se requiere una tarjeta adicional de control del O.S. (FT08. Ver Anexo E).
- Tarjeta INPUT FORMAT, especificando el formato de lectura para las nuevas variables.
- Tarjeta READ INPUT DATA, que va colocada siguiendo al primer set de tarjetas de proceso. A continuación se ponen los datos con las nuevas variables, si es que vienen en tarjetas.

El esquema del programa es el siguiente:

```
RUN NAME                opcional
GET FILE                requerida
ADD VARIABLES           requerida
INPUT MEDIUM            requerida
INPUT FORMAT            requerida
FILE NAME               opcional
VAR LABELS              opcional
VALUE LABELS            opcional
MISSING VALUES          opcional
PRINT FORMAT            requerida si alguna de las
                        nuevas variables es alfanumé-
                        rica

Tarjeta de Proceso
OPTIONS                 opcional
STATISTICS              opcional
READ INPUT DATA         requerida
:    datos
:
SAVE FILE               requerida
FINISH                  requerida
```

9.3. Eliminación de Subarchivos

El usuario puede eliminar uno o más subarchivos de un archivo por medio de la instrucción DELETE SUBFILES. Su formato es:

```
1                16
DELETE SUBFILES  subarchivo i, subarchivo k,.....
```

La tarjeta DELETE SUBFILES se pone directamente después de la tarjeta GET FILE. Por el mismo motivo que en el caso de ADD VARIABLES, los subarchivos eliminados, no pueden ser incluídos en los procesos estadísticos subsiguientes.

9.4. Agregación de Subarchivos

El usuario puede agregar casos a un archivo existente mediante la incorporación de subarchivos. Esto se hace mediante la instrucción ADD SUBFILES cuyo formato es:

```
1           16
ADD SUBFILES  lista de subarchivos que se agregan
```

El proceso es similar al de agregar variables; la posición de la tarjeta es la misma, (después de GET FILE) y también deben especificarse además las tarjetas,

a) INPUT MEDIUM *
b) INPUT FORMAT
c) # OF CASES
d) READ INPUT DATA

Desde luego los nuevos subarchivos deben contener las mismas variables y en el mismo orden del archivo original; si este último no tenía estructura de subarchivo, al agregársele subarchivos, pasa a ser el primer subarchivo con el nombre y número de casos del archivo original.

* Si se usa otro que tarjetas, debe usar tarjeta FT08. Ver Anexo E.

10. PROCEDIMIENTOS ESTADISTICOS. III
COEFICIENTES DE CORRELACION

Por medio de las tabulaciones y del test chi-cuadrado, es posible determinar si existe o no asociación entre un par de variables.

Los coeficientes de correlación tratan de cuantificar la magnitud de esta asociación, haciendo algunos supuestos sobre el tipo de relación que existe entre las variables.

Por ejemplo, el coeficiente de correlación líneal de Pearson, supone una relación de tipo lineal entre ambas variables y por esto, no es un buen indicador de la magnitud de una asociación de tipo no lineal. Esta aparente restricción se puede obviar, haciendo uso de las facilidades de transformación de variables del sistema, pudiendo así estudiar correlaciones de tipo no-lineal.(*).

10.1. Subprograma PEARSON CORR.

Este subprograma calcula el coeficiente de correlación lineal (producto-momento) de Pearson, haciendo además un test de significación sobre él.

Cuando se calculan los coeficientes de correlación entre un grupo de variables, tomadas de a pares, pueden opcionalmente ordenarse en forma matricial. Esta matriz puede ser grabada en archivos magnéticos o perforada en tarjetas, para ser usada posteriormente en otros programas del sistema o ajenas a él (se requiere en este caso una tarjeta FT09, Ver Anexo E).

La base teórica de este coeficiente, supone un nivel de medida a lo menos intervalar.

(*) Ver Anexo Estadístico

10.1.1. Tarjeta de Procedimiento

El usuario especifica los coeficientes de correlación deseados por medio de dos tipos de especificaciones. El primero consiste en dos listas de variables separadas por la palabra clave WITH. Esta especificación indica al sistema que calcule los coeficientes de correlación entre cada par de variables que se pueda formar tomando una variable de la lista que procede a la palabra WITH y la otra, de la lista que sigue a la palabra WITH.

Ejemplo:

```
1                16
PEARSON CORR  EDAD,SEXO,VAR1 TO VAR3 WITH EDUC,
              INGRESO
```

Esta tarjeta indica que se calculen e impriman los coeficientes de correlación lineal entre EDAD y EDUC, EDAD e INGRESO, SEXO y EDUC,..... hasta VAR3 e INGRESO, en total 10 coeficientes.

En el segundo tipo de especificación sólo se incluye una lista de variables y esto causa que se calculen e impriman los coeficientes de correlación entre todos los pares de variables que pueden formarse con las variables de la lista.

Ejemplo:

```
1                16
PEARSON CORR EDAD,SEXO,INGRESO,EDUC
```

Con esta especificación, se calculan e imprimen 6 coeficientes no-redundantes. (El coeficiente de correlación entre A y B es igual al entre B y A).

Sólo con este tipo de especificación se puede opcionalmente producir una matriz de correlación. (Ver Opción 4).

En este caso se perforaría o grabaría una matriz de 4x4 donde los elementos de la matriz corresponderían a los componentes dados por la siguiente tabla:

	EDAD	SEXO	INGRESO	EDUC
EDAD	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	$r_{1,4}$
SEXO	$r_{2,1}$	$r_{2,2}$	$r_{2,3}$	$r_{2,4}$
INGRESO	$r_{3,1}$	$r_{3,2}$	$r_{3,3}$	$r_{2,4}$
EDUC	$r_{4,1}$	$r_{4,2}$	$r_{4,3}$	$r_{4,4}$

Donde $r_{i,j}$ corresponde al coeficiente de correlación de la variable i con la variable j ($r_{1,2}$;EDAD con SEXO).

Se cumple que:

$r_{i,i} = 1$ para todo i

$r_{i,j} = r_{j,i}$ para todo i,j

Estas matrices pueden ser perforadas o grabadas (como imágenes de tarjetas). El formato de perforación o grabación es 8F10.7. Cada fila de la matriz comienza en una nueva tarjeta (o imagen de tarjeta), y ocupa tantas tarjetas como sea necesario. Al completarse la fila, salta a una nueva tarjeta y continúa el proceso hasta completar todas las filas de la matriz.

Se pueden combinar ambos tipos de especificación. El formato general queda entonces:

1 16

PEARSON CORR {lista de variables} [WITH {lista de variables}] / ..... / {lista de variables} [WITH {lista de variables}] /

Si se pide que se formen matrices de correlación, mediante la opción 4 (ver punto siguiente), estas se harán sólo en aquellas especificaciones en que aparezcan listas sin la palabra WITH.

Ejemplo:

```
I              16
PEARSON CORR   A,Z,B,VAR1/VAR1 TO ZETA WITH P
OPTIONS        4
```

Se perfora o graba una matriz sólo para la primera especificación, A,Z,B,VAR1.

10.1.2. Opciones

Opción 1

Esta opción causa que el programa incluya en los cálculos de los coeficientes todos los casos, independiente de si estos tienen valores declarados como MISSING.

Opción 2

Causa que los casos que contienen un MISSING VALUE sean excluídos del cálculo de todos los coeficientes, para todas las variables.

En general esta opción tiene el inconveniente de reducir considerablemente, en ciertas situaciones, el número de casos, pero por otra parte es muy útil cuando el usuario desea hacer algún análisis multivariante sobre la correlación y desea estar seguro que todos los coeficientes han sido calculados sobre el mismo conjunto de datos.

Opción 3

Se hace un test de significación de dos colas para cada coeficiente. (Ver Anexo Estadístico).

Opción 4

Produce que se perfore o grabe una matriz para todas las listas que están especificadas sin la palabra WITH.(Debe usarse tarjeta FT09).

(Todas las matrices son grabadas como imágenes de tarjeta con formato 8F10.7).

Si no se especifican opciones al programa asume:

a) Los casos con valores declarados como MISSING se excluyen sólo de los cálculos en que interviene la variable en que aparece este valor.

b) Se hace un test de significación de una cola.

c) No se graban o perforan matrices, pero las listas de correlación sin la palabra WITH son aceptadas e impresas con el formato normal.

10.1.3. Estadísticos

Los estadísticos disponibles son:

1.- Se calcula e imprime la media y desviación standard para cada variable a que se hace referencia en la tarjeta de procedimientos. Estos cálculos se hacen sobre los casos que no contienen MISSING VALUES en la variable específica a menos que se use la opción 1.

2.- Se calculan e imprimen las desviaciones de los productos cruzados y la covarianza para cada par de variables para las cuales se ha calculado el coeficiente. Estos estadísticos a diferencia de los anteriores se calculan sobre los mismos casos que se usaron para el coeficiente.

10.1.4. Limitaciones

a) Un máximo de 40 listas de coeficientes requeridos pueden aparecer en una tarjeta de procedimiento. (Las listas de coeficientes requeridos son aquellas que van entre '/', además de la primera y la última).

b) La suma de todas las variables nombradas explícita o implícitamente no puede exceder de 500 (cada vez que se nombra una variable se cuenta).

c) No más de 250 elementos pueden aparecer en el campo de especificación de la tarjeta de procedimiento. Cada nombre de variable, cada palabra WITH y cada símbolo especial cuentan como uno.

d) El número máximo de coeficientes que pueden ser calculados depende de la elección de las opciones. Si no se especifican las opciones 1 ó 2 el número máximo es (ESPACIO/24).

Si las opciones 1 ó 2 han sido especificadas el número máximo es 7.750, esto permite un tamaño máximo de la matriz de 125x125.

10.2. Subprograma NONPAR CORR

Este subprograma calcula los coeficientes de correlación por rango de Spearman y/o Kendall. El nombre del subprograma deriva que estos coeficientes son no paramétricos, es decir, no dependen de una distribución normal ni de la métrica de los intervalos. Sólo requieren escalas ordinales.

10.2.1. Tarjeta de Procedimiento

La tarjeta de procedimiento del subprograma NONPAR CORR difiere de la PEARSON CORR sólo en el campo de control, el campo de especificación es exactamente igual.

1	16
NONPAR COR	'igual tarjeta PEARSON CORR'

10.2.2. Opciones

Las opciones 1 a 4 son idénticas a las mencionadas en el punto 10.1.2. del subprograma PEARSON CORR y al no especificarlas el programa asume las mismas opciones tam

Además tiene:

Opción 5 se calculan coef. de Kendall

Opción 6 se calcula coef. de Kendall y Spearman

Si no se especifican las opciones 5 y 6 se calcula sólo el coeficiente de Spearman.

Si se especifican las opciones 4 y 6 todas las matrices de coeficientes de Kendall se graban o perforan primero que las correspondientes de Spearman.

10.2.3. Estadísticos

Los coeficientes de correlación, niveles de significación y el número de casos sobre los cuales cada coeficiente fue calculado se imprimen siempre. No hay otros estadísticos disponibles, por lo que el subprograma NONPAR CORR no utiliza la tarjeta de estedísticos.

10.2.4. Limitaciones

a) Un máximo de 100 variables pueden ser especificadas implícita o explícitamente en la tarjeta de procedimiento.

b) Un máximo de 25 listas de especificación pueden aparecer (todas las listas excepto la última terminan en un '/', esto significa que no pueden aparecer más de 24 '/').

c) A diferencia de todos los demás programas del sistema, existe un número finito de casos que pueden ser procesados. El máximo número de casos está dado por la expresión:

$$MAXCAS = \frac{(ESPACIO/2)}{2 \times NVAR + 1}$$

Donde NVAR es el número de variables usadas.

Si la opción 2 se especifica, NVAR es el número total de variables (se cuenta cada vez que aparece una variable aunque haya sido mencionada antes).

Si no se especifican las opciones 1 y 2 el denominador cambia a 2x(NVAR+2)+1.

11. TRASPASO DE ARCHIVOS Y LISTADO DE CASOS

11.1. Traspaso de Archivos

Los archivos del sistema tienen muchas ventajas para el usuario, pero tienen un inconveniente, no pueden ser leídos por otro programa de análisis que no sea un programa de SPSS. Por este motivo el sistema permite al usuario traspasar una parte o todo el archivo a tarjetas, cintas o discos, con un formato que pueda leer cualquier otro programa. Esto se hace mediante la instrucción WRITE CASES.

La instrucción WRITE CASES puede ser considerada como un procedimiento, dado que la localización de esta instrucción en la misma de la tarjeta de procesos estísticos. Sirve básicamente dos funciones:

1) Puede ser utilizada para traspasar una parte o todo un archivo del sistema a otro dispositivo.

2) Puede ser usado con un archivo del usuario de tal manera que aprovechando las facilidades de recodificación, transformación de variables y selección de casos, se cree un nuevo archivo del usuario para ser analizado por otros programas.

El formato de esta instrucción es:

```
1          16
WRITE CASES  (lista de formatos)  lista de
                                  variables
```

La lista de variables admite todos los tipos de especificaciones - posibles.

La lista de formatos es similar en estructura a la lista de formatos de la tarjeta INPUT FORMAT. Una diferencia existe en el manejo del punto decimal. En el output no se

pone punto decimal en los casos. Esto significa que el ancho del campo debe calcularse sin el punto decimal..

Ejemplo:

Rango de la variable	Formato	Rango del output
999.99 a 0.	F5.2	99999 a 0.
999.99 a -999.99	F6.2	99999 a -99999
-.3 a -1.5	F2.1	-3 a -15

La otra diferencia es que esta nueva lista de formatos admite dos elementos de formato no disponibles en la anterior. Ambos sirven la misma función, incluir en los casos una constante literal, numérica o alfanumérica, que puede servir de identificación o tener otros propósitos.

Los nuevos elementos son:

- Constante de Hollerith.

  n H c donde 'n' es un entero que indica el largo de la constante y 'c' es cualquier expresión formada con n caracteres válidos en una perforadora IBM-029.

  Ejemplo:

  6HDATOS

  constante

- Campo Literal.

  'L' donde L es una constante formada por un número cualquiera de caracteres válidos de una IBM-029 excepto el apóstrofe.

  Ejemplo:

  '25ºDATOS'

Ejemplo.

```
1                16
WRITE CASES      ('COLE=',F1.0,2X,'EDUC=',F2.0,2X,9HINGRESO'=,F7.2)
                 COLE,EDUC,INGRESO
```

'COLE=' : Escribe, perfora o graba COLE=, en las primeras 5 columnas.

F1.0 : Escribe, perfora o graba el valor de COLE en la columna 6.

2X : Se salta dos columnas, la 7 y la 8.

'EDUC=' : Escribe, perfora o graba EDUC=, en las columnas 9 a 13.

F2.0 : Escribe, perfora o graba el valor de EDUC en las columnas 14 a 15.

2X : Se salta dos columnas, la 16 y 17.

9HINGRESO'= : Escribe, perfora o graba INGRESO'= en las columnas 18 a 16.

F7.2 : Escribe, perfora o graba el valor de INGRESO en las columnas 27 a 33.

( ) : Repite el proceso para cada uno de los casos.

Este proceso requiere de una tarjeta adicional de control de O.S. FT09 (ver anexo E).

11.2. Listado de Casos

Es posible que el usuario esté interesado en examinar el contenido de algunos casos individuales especialmente cuando se han efectuado transformaciones de variables.

La tarjeta LIST CASES, permite al usuario obtener un listado de los valores de las variables que él seleccione, para los n primeros casos de cada subarchivo, donde n se especifica en la misma tarjeta de control.

Esta tarjeta no produce ningún resultado si no es usada en combinación con alguna tarjeta de procedimiento del sistema, la cual activa el procesamiento del archivo de entrada.

El formato general es:

```
1              16
LIST CASES  CASES=n(VARIABLES=lista de variables
                              ALL
```

La lista de variables admite especificaciones del tipo VARA TO VARX.

'número' : da el número de casos a listar por cada subarchivo

'lista de variables: indica las variables a listar por cada caso.

A. ESQUEMA DE USO Y DESCRIPCION DE ESTADISTICOS

Para interpretar correctamente cualquier estadístico, hay que tener muy presente la diferencia entre, describir una población (Estadística Descriptiva) e inferir acerca de una población a partir del conocimiento de una muestra (Inferencia Estadística).

Así por ejemplo, si se mide a todos los alumnos de un colegio, mayores de 15 años, y luego se calcula la altura media de ese grupo, esta medida sólo da una descripción de la muestra.

Ahora, si se quiere, a partir de ese conocimiento, inferir algo acerca de la altura de los escolares mayores de 15 años (Inferencia), deben tomarse en cuenta otros factores. Por ejemplo: ¿porqué se escogió ese colegio y no otro?, vale decir, ¿es la muestra representativa de la población?. Además debe cuantificarse el error cometido por tomar esa y sólo esa muestra.

En cuanto a la elección de los estadísticos a calcular sobre una muestra, estos están limitados por el nivel de medida de las variables que se estén considerando. Siendo el nivel más pobre que el 'nominal' (variable 'color del pelo', por ejemplo), luego el 'ordinal' (ranking de preferencias), 'intervalar' (estatura, peso). Todos los estadísticos que se pueden calcular para un nivel de medida, también pueden ser calculados para los niveles de medida superior, pero no al revés. La media aritmética, por ejemplo, sólo puede ser calculada para nivel intervalar, no teniendo sentido en niveles ordinales o nominales.

A.1. Distribuciones de Frecuencias a una Variable

Los estadísticos que se incluyen en este punto, se refieren a características de la muestra con respecto a una variable o atributo.

Los estadísticos disponibles y los subprogramas que los proporcionan están diferenciados de acuerdo al tipo de variable solamente. El usuario debe hacer la diferencia, al pedirlos, de acuerdo a los niveles de medida.

En la tabla siguiente se esquematiza la disponibilidad y el uso de los estadísticos de acuerdo al tipo de variable y al nivel de medida.

TIPO DE VARIABLES

	CONTINUA	DISCRETA	
		ALFANUMERICA	NUMERICA
SUBPROGRAMAS DISPONIBLES	CONDESCRIPTIVE	CODEBOOK (para menos de 20 cat.) MARGINALS (para más de 20 cat.)	✓ ✓ FASTMARG
ESTADISTICOS	1,2,5,6,7,8,9,10,11	No se calculan Estadísticos para Variables Alfanuméricas	1,2,3,4,5,6,7,8,9,10,11
NIVEL DE MEDIDA			
Nominal	—		4
Ordinal	—		3,4
Intervalar	1,2,5,6,7,8,9,10,11		1,2,3,4,5,6,7,8,9,10,11

Las claves de los estadísticos corresponden a:

1.- Media Aritmética	(A.1.1.3.)
2.- Error Standard de la media	(A.1.4.2.)
3.- Mediana	(A.1.1.2.)
4.- Moda	(A.1.1.1.)
5.- Desviación Standard	(A.1.2.2.)
6.- Varianza	(A.1.2.3.)
7.- Kurtosis	(A.1.3.1.)
8.- Skewness (Asimetría)	(A.1.3.2.)
9.- Rango	(A.1.2.1.)
10.- Mínimo	(A.1.3.3.)
11.- Máximo	(A.1.3.4.)

A.1.1. Estadísticos de Tendencia Central

A.1.1.1. La Moda (4)

La moda se define como el valor(es) o categoría(s) que presentan mayor frecuencia dentro de la población.

Requiere tan sólo escala nominal (por lo tanto puede ser calculada para cualquier nivel de medida) y no siempre es única.

A.1.1.2. La Mediana (3)

Se define como el valor tal que el 50% de los casos presentan valores menores que él y el otro 50% valores mayores que él.

Cuando los casos están ordenados y hay:

i) Un número N, impar, de casos, la mediana corresponde al valor del caso (N+1)/2

ii) Un número N, par, de casos, la mediana corresponde al promedio de los valores de los casos N/2 y (N/2)+1

En el caso de distribuciones continuas se define como el valor tal que la integral de -∞ a ese valor, de f(x) sea igual a 1/2.

Requiere escala ordinal a lo menos.

A.1.1.3. La Media Aritmética (1)

La media aritmética se define como la suma de valores de la variable dividida por el número total de casos. Su fórmula es:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Para el caso en que los datos estén agrupados, la fórmula queda:

$$\bar{X} = \frac{\sum f_i d_i}{N}$$

donde:

f_i es la frecuencia absoluta, y

d_i es el punto medio de la categoría i

Requiere a lo menos escala intervalar.

A.1.2. Estadísticos de Dispersión

Todos ellos requieren niveles de medida a lo menos intervalar.

A.1.2.1. Rango (9)

Se define como la diferencia entre el mayor valor y el menor valor que toma la variable.

Este estadístico da una idea de la concentración de los datos.

A.1.2.2. Desviación Standard (5)

La desviación standard se define como la raiz cuadrada del promedio de las desviaciones cuadráticas con respecto a la media. Su fórmula es:

$$\left[\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}\right]^{1/2}$$

para datos no agrupados

$$\left[\frac{\sum_{i=1}^{N} f_i X_i^2}{N}\right]$$

para datos agrupados

donde:

$$X_i = d_i - \bar{X}$$

A.1.2.3. Varianza (6)

La varianza se define como el cuadrado de la desviación standard.

Ambos estadísticos dan una medida de la dispersión con respecto a la media.

A.1.3. Otros Estadísticos e Indicadores

A.1.3.1. Kurtosis (7)

La Kurtosis es un estadístico que mide desviaciones de la distribución de los datos con respecto a una distribución normal. Su fórmula es:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{S}\right)^4}{N} - 3$$

Valores positivos indican que la distribución es más puntiaguda que la distribución normal y valores negativos, que es más achatada.

A.1.3.2. Skewness (Asimetría) (8)

Una distribución es asimétrica cuando hay un número considerable de casos extremos más en un lado de la distribución que en el otro. Su fórmula es:

$$\text{Skewness} = \frac{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{S}\right)^3}{N}$$

Valores positivos indican que la distribución es asimétrica a la derecha. Cuando el resultado es negativo es asimétrica a la izquierda.

A.1.3.3. Mínimo (IO)

Se define como el menor valor que toma la variable.

A.1.3.4. Máximo

Se define como el mayor valor que toma la variable.

A.1.4. Nociones sobre Teoría de Muestreo

Se puede decir que, en este caso, el problema característico es obtener estimaciones de los parámetros de una población a partir de información disponible acerca de una muestra de esa población, tratando de minimizar y cuantificar el error de la estimación.

Aquí, sólo se abordará el problema de estimar la media poblacional a partir de la media muestral y estimar el parámetro 'p' de una población binomial a partir del p_m muestral.

FILE DATOS (CREATION DATE = 03/13/73) FICTICIO DE CUESTIONARIO

CROSSTABULATION OF SEXO BY INGRESO

SEXO \ INGRESO	MENOS DE 5 MIL	ENTRE 5 Y 15	MAS DE 15	ROW TOTAL
0.00	13	18	5	36
	36.1	50.0	13.9	60.0
	48.1	69.2	71.4	
	21.7	30.0	8.3	
1.00	14	8	2	24
	58.3	33.3	8.3	40.0
	51.9	30.8	28.6	
	23.3	13.3	3.3	
COLUMN TOTAL	27	26	7	60
	45.0	43.3	11.7	100.0

CHI SQUARE = 2.88427 WITH 2 DEGREES OF FREEDOM

CRAMER'S V = 0.21929

CONTINGENCY COEFFICIENT = 0.21416

GOODMAN'S TAU B = -0.20356

GOODMAN'S TAU C = -0.21778

GAMMA = -0.37121

SOMER'S D = -0.18267

A.1.4.1. Distribuciones Muestrales

Si se consideran todas las muestras de tamaño n, que pueden ser tomadas de una población dada, entonces para cada muestra, se puede calcular un estadístico (media, desviación standard, moda, etc.), que variará de una muestra a otra, en forma aleatoria.

La distribución del estadístico, obtenida de las muestras, es llamada 'distribución muestral'.

La desviación standard de la distribución muestral se denomina 'error standard'.

A.1.4.2. Distribución de las Medias Muestrales

Si todas las muestras de tamaño n, son sacadas, sin reemplazo, desde una población de tamaño $N > n$ y si llamamos $\mu_{\bar{X}}$ y $\sigma_{\bar{X}}$ a la media y desviación standard de la distribución de la media muestral y μ y σ a la media y desviación standard de la población, entonces

$$\mu_{\bar{X}} = \mu \quad \text{y} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Si la población es infinita o las muestras de tamaño n son sacadas con reemplazo(*) desde una población finita, entonces

$$\mu_{\bar{X}} = \mu \quad \text{y} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Cuando la población sigue una ley normal, la distribución de las medias muestrales es también normal.

(*) Es decir, se puede sacar infinitas muestras de cualquier tamaño.

Para valores grandes de n($n \geq 30$), y poblaciones infinitas o muestras con reemplazo de poblaciones finitas, la distribución de las medias muestrales se aproxima a una distribución normal, con media $\mu_{\bar{X}}$ y desviación standard $\sigma_{\bar{X}}$, aunque la población no siga una ley normal.

A.1.4.3. Distribución de Proporciones Muestrales

Si se tiene una población infinita, en que la ocurrencia de un suceso (llamado éxito), tiene probabilidad p y la no ocurrencia de ese suceso tiene probabilidad q=1-p, entonces al considerar todas las posibles muestras de tamaño n (son infinitas) y determinar la proporción de exitos, se obtendría la distribución muestral de las proporciones p_m, cuya media μ_p y desviación standard σ_p están dadas por:

$$\mu_p = p \quad y \quad \sigma_p = \sqrt{\frac{p\,q}{n}}$$

Para valores grandes de n ($n \geq 30$) la distribución muestral de los p_m se puede considerar normal.

A.1.4.4. Estimación de Parámetros

Sea μ_s y σ_s la media y desviación de la distribución muestral de un estadístico S. Entonces si la distribución muestral se puede considerar normal, se espera que el valor del estadístico S, correspondiente a una muestra cualquiera esté, por ejemplo, entre los valores $\mu_s - \sigma_s$ y $\mu_s + \sigma_s$ con una probabilidad de .682. De ahí se tiene que:

$$\left.\begin{array}{l} \mu_s - \sigma_s \leq S \Longrightarrow \mu_s \leq S + \sigma_s \\ S \leq \mu_s + \sigma_s \Longrightarrow \mu_s \geq S - \sigma_s \end{array}\right\} \Longrightarrow$$

$S - \sigma_s \leq \mu_s \leq S + \sigma_s$ con una probabilidad de .682

Más general:

$$S - K\sigma_s \leq \mu_s \leq S + K\sigma_s$$

con una probabilidad dada por la siguiente tabla:

TABLA A.1.

Prob.	.997	.99	.98	.96	.95.4	.95	.90	.80	.682	.5
K	3.0	2.58	2.33	2.05	2.00	1.96	1.64	1.28	1.0	.67

- Así para estimar la media de una población a partir de una media muestral μ_m calculada sobre una muestra de tamaño n=50, podemos fijar un intervalo para el valor μ, tal que μ esté en ese intervalo con una probabilidad dada, ya que n=50 asegura que la distribución muestral se puede considerar normal. Así, por ejemplo:

$$\mu_m - 2\sigma_{\bar{X}} \leq \mu_{\bar{X}} \leq \mu_m + 2\sigma_{\bar{X}} \quad \text{con probabilidad .954}$$

Dado que $\mu_{\bar{X}} = \mu$ tenemos que:

$$\mu_m - 2\sigma_{\bar{X}} \leq \mu \leq \mu_m + 2\sigma_{\bar{X}} \quad \text{con probabilidad .954}$$

Ahora bien, $\sigma_{\bar{X}}$ está dado por:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Como σ no se conoce (en la mayoría de los casos) se usa una estimación de σ^2 dada por:

$$\hat{S}^2 = \frac{n}{n-1} S^2 \quad \text{donde } S^2 \text{ es la varianza de la muestra}$$

Asi finalmente se tiene que, μ es tal que:

$$\mu_m - 2\frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \mu_m + 2\frac{\hat{S}}{\sqrt{n}} \quad . \quad (*)$$

con probabilidad .954

Mas general, si la distribución puede considerarse normal, μ , con probabilidad dada por la tabla A.1., es tal que:

$$\mu_m - K\frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \mu_m + K\frac{\hat{S}}{\sqrt{n}} \quad (*)$$

-Para estimar el p de una distribución binomial tenemos que, para $n \geq 30$, se cumple:

$$P_m - K\sqrt{\frac{pq}{n}} \leq p \leq P_m - K\sqrt{\frac{pq}{n}}$$

Dado que p no se conoce y por lo tanto tampoco q, se usa la siguiente aproximación.

$$P_m - K\sqrt{\frac{P_m(1-p_m)}{n}} \leq p \leq p_m - K\sqrt{\frac{p_m(1-p_m)}{n}} \quad (*)$$

(*) En rigor, cuando se usan estimaciones del error standard, K debe calcularse según una distribución t-Studen y no según una normal como se ha hecho, pero para $n \geq 30$ el error que se comete no es significativo.

A.2.- Distribuciones de frecuencia a dos o más variables.

Hasta este momento, los elementos han sido medidos con respecto a solo una variable. Ahora los ca-

sos contienen información acerca de dos o más atributos y el problema consiste en determinar si existe asociación entre estas variables. Se dice que existe asociación cuando la información acerca de una de las variables permite predecir el valor de otra variable para el mismo caso. Por ejemplo, suponga que se tiene información acerca de la altura de un grupo de padres ¿Ayuda esta información a predecir la altura de sus hijos?. Si no ayuda, entonces no estarán asociados.

Los subprogramas de tabulaciones traen diversos test de asociación y coeficientes de correlación (los que se discuten en el siguiente capítulo del Anexo). De los test, los más importantes se discuten en detalle y los menos usados, sólo someramente.

A.2.1. Test de Asociación

Estos test dependen fundamentalmente del tamaño de la muestra y del nivel de medida. Los que aquí se describen requieren tan sólo escalas nominales.

A.2.1.1. Test Exacto de Fisher

El test exacto de Fisher es una técnica noparamétrica sumamente podetosa para analizar tablas de 2 x 2 cuando la muestra es pequeña.

Este test sirve para decidir si las variables estan asociadas o no. Considerando los totales marginales como fijos y suponiendo que no hay asociación, calcula la probabilidad exacta (de ahí su nombre) de que en una muestra aparezca esa distribución o una que aún de más evidencia de asociación. Si esta probabilidad es más pequeña que el nivel de significación que se desea, se rechaza la hipótesis de no asociación.

La probabilidad de una distribución particular se deduce a partir de la distribución hipergeométrica y está dada por:

$$p_i = \frac{(A+B)!\ (C+D)!\ (B+D)!\ (A+C)!}{N!\ A!\ B!\ C!\ D!}$$

donde A, B, C, D y N corresponden a las frecuencias en las celdas y al total de casos, como se muestra en la figura

		VARIABLE 2		
		+	-	
Variable 1	0	A	B	A+B
	1	C	D	C+D
		A+C	B+D	N

Ejemplo: Se tiene una muestra de 12 individuos y se quiere decidir en base a esta muestra si existe o no asociación entre el color del pelo y el color de los ojos.

Los datos tabulados son:

	Azules	Cafes	
Rubio	4	1	5
Castaño	1	6	7
	5	7	12

Hipótesis: No hay asociación entre el color del pelo y el color de los ojos.

Test: Fisher exacto

Nivel de Significación: 0.05

Se calcula el P_i para la tabla original y para todas las distribuciones que den aún más evidencia de asociación, teniendo los totales marginales fijos. Se suman todos los p_i y se obtiene la probabilidad de ocurrencia, bajo la hipótesis de no asociación de la distribución dada por la muestra o de una distribución más extrema.

Si la probabilidad $p=\sum p_i$, es menor que el nivel de significación usado, se rechaza la hipótesis de no asociación.

En el ejemplo, la única distribución más extrema es:

	Azules	Cafes	
Rubio	5	0	5
Castaño	0	7	7
	5	7	12

$$p_2 = \frac{5!\ 7!\ 5!\ 7!}{12!\ 5!\ 7!\ 0!\ 0!} = 0.001$$

$$p = p_1 + p_2$$

$$p_1 = \frac{5!\ 7!\ 7!\ 5!}{12!\ 4!\ 1!\ 1!\ 6!} = 0.044$$

$$p = 0.045$$

Luego con un nivel de significación del 0.05 podemos rechazar la hipótesis de que no existe asociación entre las variables. Es decir la probabilidad de equivocarse al rechazar la hipótesis es menor que 0.05, (el máximo, que en este ejemplo se permite).

A.2.1.2. Test de Chi-cuadrado

Cuando se tiene una tabla de mayores dimensiones o una tabla de 2 x 2 pero con gran número de casos se usa el test de Chi-cuadrado. Este test mide la significación de las diferencias entre las frecuencias observadas (O_i) y las frecuencias esperadas bajo la hipótesis (E_i).

La idea es que bajo la hipótesis de no asociación las frecuencias relativas de cada fila o columna no tengan diferencias significativas con respecto a las distribuciones marginales.

El test calcula una función de las diferencias entre las frecuencias esperadas y las observadas, dada por:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde r es el número de filas y c el número de columnas y si la hipótesis es de no asociación se tiene que las frecuencias esperadas estan dadas por

$$E_{ij} = \frac{f_i\ f_{ic}}{N}$$

f_i : frecuencia marginal de la columna j

f_{ic} : frecuencia marginal de la fila i

Entonces el estadístico χ^2, así calculado, se distribuye como chi-cuadrado con (r-1) (c-1) grados de libertad. Esta distribución está tabulada y dá la probabilidad de que χ^2 sea mayor que el valor de chi-cuadrado.

Ejemplo: Suponga que se quiere decidir si existe asociación entre el grado de alcoholismo y el grado de habilidad manual en un grupo de obreros.

Los datos están tabulados de la siguiente manera:

Habilidad Manual

Grado Alcoholismo	E	B	R	M	
Alto	1	3	15	20	39
Medio	11	13	25	23	72
Bajo	18	24	20	27	89
	30	40	60	70	

La hipótesis es: No existe asociación entre la habilidad manual y el grado de alcoholismo.

Test de Asociación: Chi-cuadrado

Nivel de significación: 0.05

$$\chi^2 = \frac{(1-5.85)^2}{5.85} + \frac{(3-7.8)^2}{7.8} + \frac{(3-7.8)}{7.8} + \frac{(20-13.65)^2}{13.65}$$

$$+ \frac{(11-10.8)^2}{10.8} + \frac{(13-14.4)^2}{14.4} + \ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots + \frac{(27-31.16)^2}{31.15} = 17.73$$

Se ve en la tabla de chi-cuadrado, que para un valor de χ^2 de 17.73 con 6 grados de libertad, se tiene una probabilidad p=0.01 que $\chi^2 \geq 17.73$.

Dado que p=0.01 es menor que el nivel de significación usado, se rechaza la hipótesis de no asociación.

Este test puede ser usado con cualquier nivel de medida siempre que las frecuencias esperadas en cada celda no sean demasiado pequeñas. Se recomienda que para χ^2 con más de 1 grado de libertad, no más del 20% de las celdas tengan frecuencias esperadas menores que 5 y ninguna celda tenga frecuencia esperada cero.

Para el caso de tablas de 2 x 2 se tiene que:

a) Cuando $N>40$ se usa chi-cuadrado con la corrección de Yates.

b) Cuando $21 \leq N \leq 40$, se puede usar chi-cuadrado si todas las frecuencias esperadas son de 5 o más. Si no se debe usar Fisher.

c) Cuando $N < 21$ se debe usar Fisher en todos los casos.

El sistema no hace la diferencia del punto b, es decir en tablas de 2 x 2 si hay menos de 21 casos usa Fisher, si no chi-cuadrado. Por lo tanto el usuario debe tomar algunas precauciones al interpretar el test ya que puede suceder que este, no tenga significación.

A.2.2. Coeficientes de Asociación Basados en Chi-cuadrado

A.2.2.1. Phi.

Phi es una corrección a X^2 por el hecho de que este último es directamente proporcional a N. Su formula es:

$$\text{Phi } (\emptyset) = \frac{\chi^2}{N}$$

Para una tabla de 2 x 2 su valor varía de 0, cuando no existe asociación entre las dos variables, a 1, cuando la asociación es perfecta.

A.2.2.2.- V. de Cramer

Cuando Phi se calcula para tablas que no son de 2 x 2, no tiene límite superior. Por esto el V de Cramer es usado para ajustar phi, ya sea por el número de columnas o filas de la tabla, dependiendo de cual es el menor. Su formula es:

$$V = \left\{ \frac{\phi^2}{\text{Min}\left[(r-1),\ (c-1)\right]} \right\}^{1/2}$$

Su valor varía entre 0 y 1, independiente del tamaño de la tabla.

A.2.2.3.- Coeficiente de Contingencia

Este coeficiente también surge de la necesidad de ajustar χ^2 por el hecho de ser proporcional a N. Su formula es:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Fue propuesto porque se puede mostrar que si una distribución normal bivariada con coeficiente de correlación ρ (Pearson) se clasifica en una tabla de contingencia, entonces $C^2 \rightarrow \rho^2$ cuando el número de categorías en la tabla tiende a infinito. Para r y c finitos, sin embargo, el coeficiente tiene sus limitaciones. Se hace cero, como es normal, cuando hay total independencia, pero en general C no puede ser nunca uno y su limite superior varía según el número de filas y columnas de la tabla.

No se puede pretender que cualquier coeficiente basado en χ^2 pueda mostrarse como una medida satisfactoria de asociación, principalmente por el hecho de que sus valores no tienen una interpretación clara de tipo probabilístico.

En el próximo punto de este Anexo se analizan otros coeficientes de asociación para escalas nominales y ordinales que tienen una más clara interpretación.

A.2.3.-Asociación entre Escalas Ordinales y Nominales:

Con estos niveles de medida, los coefi - cientes que pueden definirse no siempre tienen una interpretación que indique claramente de que manera ayuda el conocimiento de una variable a predecir el valor de la otra.

A.2.3.1.- Coeficiente de Asociación de Goodman-Kruskal (GAMMA)

La idea de este coeficiente es que a partir de él, se pueda tener información acerca de la dirección de cambio de una variable conociendo la variación de la otra.

Se entenderá mejor el coeficiente a través del siguiente ejemplo:

Suponga que dos jueces x e y han calificado a un grupo de 4 alumnos, de acuerdo al grado de excelencia en un trabajo. El resultado esta dado por la tabla siguiente:

	Jueces	
Alumnos	X	Y
A	3	3
B	4	1
C	2	4
D	1	2

Reordenando los alumnos de manera que las calificaciones del Juez esten en el orden normal. Se tendría entonces:

	X	Y
D	1	2
C	2	4
A	3	3
B	4	1

Lo que pretende este coeficiente es dar una medida del grado de acuerdo o desacuerdo que existe entre ambos jueces.

Para ello se determina cuantos pares de calificaciones del Juez Y están en el orden natural y cuantas no

lo están.

Así los pares (2, 4) y (2, 3), están en orden natural, los pares (2, 1), (4, 1), (3, 1) y (4, 3) no están en orden natural.

Al número de pares en orden natural se les llama número de concordancias ,porque concuerda con el orden de las calificaciones del Juez Y . Se denota fc.

Al número de pares en orden invertido se le llama número de inversiones y se denota por fi.

El coeficiente γ se define como la diferencia entre concordancia e inversiones, relativa a la suma de ambos.

$$\gamma = \frac{fc - fi}{fc + fi}$$

Este coeficiente varía entre 1 y -1. Toma el valor 1 cuando el número de inversiones es cero, es decir cuando hay acuerdo total entre los dos jueces. El valor -1 corresponde al caso en que el número de concordancia es cero, es decir hay total desacuerdo entre ambos jueces.

En el ejemplo:

$$\gamma = \frac{2 - 4}{2 + 4} = -0.33$$

Este valor se interpreta diciendo que hay un 33% de más de desacuerdo entre ambos jueces que de acuerdo.

La distribución muestral de γ es aproximadamente normal para $N \geq 8$ con:

$$\mu_\gamma = 0 \quad y \quad \sigma_\gamma = \sqrt{2\ (2N + 5) / 9N\ (N + 1)}$$

A.2.1.2.- Coeficientes de Kendall Tau B Tau C.

- Tau B

Su definición e interpretación coincide con el coeficiente r en el caso en que no hay empates (En el ejemplo anterior habría un empate si el Juez x o y hubiera dado la misma nota a dos alumnos).

En el caso en que hay empates la formula de Tau B queda como sigue:

$$\text{Tau B} = \frac{fc - fi}{\sqrt{\left(\frac{N(N-1)}{2} - Ex\right)\left(\frac{N(N-1)}{2} - Ey\right)}}$$

Donde Ex es igual al número de observaciones empatadas en la variable X. E_y está definido analogamente.

-Tau C

Esta dado por:

$$\text{Tau C} = \frac{2m\,(fi-fi)}{N^2\,(m-1)}$$

donde m= mínimo {N°de filas, N°de columnas}

A.2.3.3.- Coeficiente de Predicción de Guttman (LAMBDA)

Se utiliza, para detectar asociación entre escalas nominales.

Dependiendo de si se supone que una de las variables en particular es dependiente o si no se supone, se usan dos formas distintas del coeficiente.

- Lambda Asimétrico

En este caso el coeficiente está dado

$$\lambda a = \frac{Rd}{E}$$

donde Rd=reducción del error de la variable dependiente
E =error total

(Cabe hacer notar la similitud con una de las interpretaciones del coeficiente de correlación lineal dada por el cuociente entre la reducción del error y el error total).

Suponga que, por cualquier motivo, cierta empresa ha decidido hacer un estudio de la relación que existe entre el color del pelo y la marca de cigarrillos que contienen las personas.

Se encuestan a 55 fumadores, obteniéndose la siguiente tabla:

Color \ CIG.	Lucky	Life	Hilton	Total	
Oscuro	10	5	3	18	(N_1)
Castaño	6	4	8	18	(N_2)
Rubio	2	7	10	19	(N_3)
Total	18	16	21	55	(N)
			Fd.		

Si se quiere predecir la marca de cigarrillos que consume una persona, sin importar el color del pelo, naturalmente se recurrirá a la clase modal, es decir HILTON.

El error que se comete es 55-21=N-Fd.

Si se toma en cuenta el color del pelo, se hará la predicción según la clase modal de la fila. Así, si se sabe que el pelo es OSCURO se dirá que la persona fuma LUCKY y se cometerán errores en $18-10=N_i-F_i$ casos.

Así en total, al considerar la variable "color del pelo" para predecir la variable "tipo de cigarrillos", el error que se comete está dado por

$$(N_1-f_1) + (N_2-f_2) + (N_3-f_3) = N-\sum f_i$$

La reducción del error es la diferencia entre el error que se comete al considerar la segunda variable. Esto es:

$$(N-fd) - (N-\sum fi) = \sum fi - Fd.$$

De aquí resulta que el coeficiente tiene la siguiente forma:

$$\lambda_a = (\sum fi - Fd) / (N - Fd)$$

Del Ejemplo:

$f_1 = 10$, $f_2 = 8$, $f_3 = 10$ $\quad fi = 28$

$Fd = 21$, $N = 55$

$$\lambda = (28-21) / (55-21) \approx 0.2$$

Esto se interpreta diciendo que el error de la predicción en el consumo de cigarrillos disminuye en 20% al considerar el color del pelo.

- Lambda Simétrico

En este caso no se supone que una variable en particular sea dependiente de la otra.

Se define como

$$\lambda = R_t / E_t$$

donde Rt es la reducción del error en ambas variables y Et es el error original en ambas variables.

La fórmula en este caso es:

$$\lambda = \frac{\sum fi_1 + \sum fi_2 - F_1 - F_2}{N-F_1 + N-F_2}$$

donde f_{ij} es la frecuencia máxima encontrada dentro de cada subclase de la variable J (J= 1, 2)

F_j es la frecuencia modal de la variable J (J=1, 2)

A.2.3.4.- Coeficiente D de Sommer (Sommer's D)

El coeficiente D de Sommer es similar al coeficiente Gamma, pero concidera los empates como información válida.

Su fórmula es:

$$D = \frac{2\,(P - Q)}{1/2\left[(N^2 - C_j^2) + (N^2 - R_j^2)\right]}$$

C_j : Es el total de la columna j

R_j : Es el total de la fila j

A.2.3.5.- Coeficiente de Incertidumbre.

Este coeficiente está basado en la Teoría de Información y de una medida de la información que dá el conocimiento de la variable independiente acerca de la variable dependiente.

En términos propios de la Teoría de Información, da el grado de reducción de la Entropía del sistema.

El coeficiente de incertidumbre asimétrico está dado por:

$$U_a = \frac{U\,(Y) - U\,(Y/X)}{U\,(Y)}$$

Donde U(Y) es la entropía de Y , que está dada por

$$U\,(Y) = -\sum_j p\,(Y_j) \log p\,(Y_j)$$

Donde $p(Y_j)$ es la probabilidad de la categoría y en la distribución de Y.

$$U\,(Y/X) = -\sum_k \sum_j p\,(Y_j;\, X_k) \log p\,(Y_j/X_k)$$

El coeficiente de incertidumbre simétrico está dado por

$$U_s = \frac{U\,(Y) + U\,(X) - U\,(Y, X)}{U\,(Y) + U\,(X)}$$

Donde U (Y,X) está dado por

$$U\,(X,X) = -\sum_j \sum_k p\,(Y_j, X_k) \log p\,(Y_j\, X_k)$$

Ambos coeficientes varían entre 0 y 1

A.3.- Coeficiente de Correlación.

La idea fundamental de la correlación deriva del deseo de poder resumir el grado de asociación entre variables estadísticas en un coeficiente.

Es interesante notar que no importa cuan fuerte sea una relación estadística, acusada por un coeficiente, esto no implica una relación de tipo causal, la causalidad debe venir de fuera de la estadística. Así, por ejemplo, puede establecerse que existe una gran correlación entre la tasa de nacimientos y la llegada de las cigueñas a Paris, pero este es un caso donde obviamente la correlación no es causal.

El problema de la asociación entre variables estadísticas puede plantearse de manera simple, diciendo que esta existe, cuando la información que se tiene sobre una variable permite predecir el comportamiento de la otra. Por ejemplo, si el conocimiento del peso de un individuo nos ayuda a predecir su altura, diremos que ambas variables están correlacionadas.

Dependiendo del nivel de medida de las variables, se utilizarán coeficientes apropiados. Estos diferentes coeficientes tienen el inconveniente de no ser comparables entre si. Así no son comparables coeficientes distintos calculados sobre una misma muestra, ni tampoco el mismo coeficiente calculado sobre muestras distintas. Lo que si es comparable en ambos casos es el nivel de significación que, para estos coeficientes, dan los test de significación.

A.3.1.- Coeficiente de Correlación Lineal de Pearson.

Se han medido las características X e Y en un grupo de individuos de una cierta población, obteniéndose la distribución conjunta de X e Y.

Si no se tuviera información sobre X, pa-

ra predecir el valor de la variable Y para un individuo en particular (Y_i), se utilizaría el valor de la media $\bar{Y}$. El error cuadrático medio que se comete, esta dado por

$$E = \sum_i (Y_i - \bar{Y})^2 / N$$

Si se supone una relación de tipo lineal entre la variable X e Y, dada por

$$Y'_i = a X_i + b$$

donde $Y'_i = Y_i + E_i$ con E_i un error aleatorio, se espera que el error cometido al predecir Y_i con Y'_i sea menor que el cometido al predecir con $\bar{Y}$, es decir

$$\sum_i (Y_i - Y'_i)^2 / N \leq \sum_i (Y_i - \bar{Y})^2 / N$$

Lo que trata de indicar el coeficiente de correlación lineal para predecir Y_i, entregando el porcentaje de disminución del error cometido en este caso con respecto al error original E

De lo anterior se tiene que

$$\rho^2 = \frac{\sum (Y_i - \bar{Y})^2 / N - \sum (Y_i - Y'_i)^2 / N}{\sum (Y_i - \bar{Y})^2 / N}$$

Si efectivamente la suposición de relación o dependencia lineal es buena, se tendrá que el error $1/N \sum (Y_i - Y'_i)^2$ será pequeño, lo que hace que $\rho^2 = 1$. Si por el contrario la suposición de linealidad es desacertada, este último error será comparable al primero, con lo cual se tiene $\rho^2 = 0$.

De la misma definición de ρ^2, surge su interpretación; Por ejemplo, si $\rho^2 = .65$ se dirá que el error que se comete al predecir Y con Y', Si $\rho^2 = .00$ significa que con la suposición de relación lineal X no ayuda a predecir Y. (Esto no significa necesariamente que las variables sean independientes, ya que puede tenerse una relación del tipo $X_i Y_i = 1$ por ejemplo, la que no se reflejaría en el coeficiente de correlación lineal).

Tambien puede decirse que un porcentaje igual a ρ^2 de la varianza de Y es explicado por la varianza de X.

Se puede demostrar que si la ecuación le recta $Y_i = a X_i + b$, correpponde a la recta de mínimos cuadrados, se tiene que:

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - Y_i')^2 + \sum (Y_i' - \bar{Y}^2)$$

de donde:

$$\rho^2 = \frac{\sum (Y_i' - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

lo que puede interpretarse diciendo que ρ^2 es el cuociente entre la variación explicada y la variación total.

Reemplazando Y_i por $aX_i + b$ donde a y b se obtienen de la expresión para recta de mínimos cuadrados se llega a:

$$\rho^2 = \frac{\left[\sum (X_i - \bar{X})\ (Y_i - \bar{Y})\right]^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}$$

que es la más conocida de ρ^2.

Lo que usualmente se calcula es una estimación, r, del coeficiente ρ (ya que generalmente los cálculos se realizan sobre una muestra de la población). El signo de r corresponde al signo de la pendiente de la recta que representa la relación lineal.

Existen dos test de significación de acuerdo a la hipótesis que se haga sobre ρ

- Hipótesis : $\rho = 0$

Se usa el hecho de que el estadístico

$$t = r \sqrt{N - 2} \ / \sqrt{1 - r^2}$$

sigue una distribución t de Student con $N - 2$ grados de libertad.

Así por ejemplo, si en una muestra de 18 individuos se observa un valor r = 0.2 Se tiene que:

$$t = 0.2 \times \sqrt{18 - 2} \; / \; \sqrt{1 - 0.04} = 0.82$$

Según la tabla de la distribución t de Student, la probabilidad de que, si la hipótesis es correcta, se observe en una muestra un valor de t de 0.82 o mayor es más de 0.20. Esto quiere decir que la probabilidad de equivocarse al rechazar la hipótesis es mayor que 0.20

Si se quiere tener un nivel de significación menor o igual a 0.20, debe aceptarse la hipótesis. Es decir si se quiere tener una probabilidad menor o igual a 0.20 de equivocarse al rechazar la hipótesis.

Si para la misma muestra se obtiene un valor r = .7 (para otro par de variables), se tendría t = 3.9. La probabilidad de observar en una muestra un valor t de 3.9 o mayor es menor que .005. Luego, con una probabilidad de equivocarse menor que .005, se puede rechazar, en este caso, la hipótesis $\rho = 0$.

- Hipótesis $\rho = \rho_0 \neq 0 \wedge \rho_0 \neq 1$

Se usa el hecho de que el estadístico

$$Z = \left[\log_e \left\{(1+r) \; / \; (1-r)\right\}\right] / 2$$

tiene una distribución aproximadamente normal, con media μ y desviación standard σ_z dada por:

$$\mu_z = \left[\log \left\{(1 + \rho_0) \; / \; (1-\rho_0)\right\}\right] /2 \quad \text{y} \quad \sigma_z = 1/\sqrt{N-3}$$

luego dado un valor de Z, obtenido de una muestra de tamaño N tenemos que:

$$\mu_z - K\sigma_z \leq z \leq \mu_z + K\sigma_z$$

con una probabilidad dada por el valor de K, según la tabla A.1.

Ej. De una muestra de 103 individuos, se obtuvo un valor de r de .750, se quiere tener una estimación de ρ_0 a partir del valor r ob-

tenido. Se tiene que

$$Z = \left[\log (1.75/0.25)\right] /2 = 1/2 \log 7 = .42$$

$$\sigma_Z = 0.1$$

$$\mu_Z - K*0.1 \leq Z \leq \mu_Z + K*0.1$$

$$\mu_Z \leq Z + 0.1.K$$

$$\mu_Z \geq Z - 0.1.K$$

Si se quiere tener una probabilidad de ,95 se usa K = 1.96

De ahi

$$\mu_Z \leq 0.42 + 0.196 = 0.616$$

$$\mu_Z \geq 0.42 - 0.196 = 0.224$$

$$1/2 \log \left[(\rho_o+1)/(1-\rho_o) \right] \leq 0.616$$

$$\log \left[(\rho_o+1)/(1-\rho_o) \right] \leq 1.232$$

$$(\rho_o+1)/(1-\rho_o) \leq 17.0$$

$$1 + \rho_o \leq 17.0 \ (1-\rho_o)$$

$$18.\rho_o \leq 16.0$$

$$\rho_o \leq 16.0/18.0 = .89$$

de la misma forma se obtiene

$$\rho_o \geq .47$$

Asi se puede decir que, con una probabilidad de .95, es tal que

$$.47 \leq \rho_o \leq .89$$

A.3.2.- Coeficientes de Correlación Noparamétricos Spearman y Kendall.

Estos coeficientes de correlación tienen la particularidad de no depender de una distribución normal de la población o de las propiedades métricas de las escalas inter valores.

La principal diferencia entre el coeficiente r_s de Spearman y el coeficiente tau de Kendall, parece ser

que este último es más significativo cuando los datos contienen un gran número de empates. Por otra parte r_s parece ser una mejor aproximación al coeficiente de correlación producto momento (r), cuando los datos son más o menos continuos (no presentan un gran número de empates en cada rango).

Ambos coeficientes varian entre 1.0 y - 1.0 Además ambos requieren que las variables sean numericas y el nivel de medida sea ordinal.

A.3.2.1.- Coeficiente de Correlación de Spearman.

Se define como:

$$r_s = \frac{T_x + T_y - \sum d_i^2}{2\sqrt{T_x T_y}}$$

donde: T_x (T_y) está definido por

$$\frac{N(N^2 - 1) - \sum R(R^2 - 1)}{12}$$

donde R es el número de empates en rango dado para la variable X (Y).

La significación de este coeficiente, dada la hipótesis $\rho_s = 0$, puede determinarse usando el hecho de que el estadístico.

$$t_s = r_s (N-2) / \sqrt{1 - r_s^2}$$

sigue una distribución t de Student con N-2 grados de libertad.

A.3.2.2.- Coeficiente de Correlación de Kendall.

El coeficiente de correlación de Kendall está dado por:

$$\text{Tau} = \frac{S}{\sqrt{1/2\ N\ (N-1) - T_x}\ \sqrt{1/2\ N\ (N-1) - T_y}}$$

donde S = número de concordancias - número de inversiones.

Y $T_x = 1/2 \sum (t-1)$, donde t es el número de empates en cada rango de la variable X.

T_y se define en forma análoga.

La significación de este coeficiente, dada la hipótesis $\tau = 0$, puede determinarse usando el hecho de que Tau sigue una distribución normal con media 0 y desviación standard dada por:

$$\sigma_\tau = \sqrt{\frac{4N + 10}{9N(N-1)}}$$

donde N es el tamaño de la muestra

C. LIMITACIONES Y CONVENCIONES DE SUBPROGRAMAS ESTADISTICOS. RESUMEN

C.1. CONDESCRIPTIVE

{ ALL
{ lista de variables

Tipo de variables : Numéricas Continuas
Tipo de Lista : Acepta especificaciones VARA TO VARX, admite hasta 250 variables. (VARA TO VARX cuenta sólo como 3)

C.2. CODEBOOK

{ ALL
{ lista de variables

Tipo de variable : Alfanumérica, numérica discreta (no más de 20 categorías)
Tipo de Lista : Acepta especificaciones VARA TO VARX, admite hasta 250 variables. (VARA TO VARX cuenta sólo como 3)

Número Máximo de Variables a Procesar : Está dado por:

$$MAXVAR = \frac{(ESPACIO/4)-(*xNCAT)-15}{4+2xNCAT}$$

NCAT: es el número de categorías en la variable que tiene más.
ESPACIO: espacio asignado en memoria, es modificable y su valor usual es 50.000.-

C.3. MARGINALS { ALL / lista de variables

Tipo de variables : Alfanumérica o Numérica discreta (se usa normalmente con más de 20 categorías)

Tipo de Lista : Acepta especificaciones VARA TO VARX y admite hasta 250 variables (VARA TO VARX cuentan como 3)

N°de Variables a procesar : Está dado por:

$$MAXVAR = \frac{(ESPACIO/4) - 3(NCAT)}{(2 \times NCAT) + 2}$$

C.4. FASTMARG : { Lista de variables (mínimo, máximo)/.....

Tipo de Variables : Numérica entera

Tipo de Lista : Acepta especificaciones VARA TO VARX.

Espacio Requerido : El espacio mínimo requerido está dado por:

ESPACIO = 4 x SOMVAL + 20 MAXROT

MAXROT es el número máximo de categorías que puede tener una variable (si es una opción 2 o 5 MAXROT = 0).

SOMVAL es la suma del número de categorías de todas las variables

C.5. CROSSTABS : {Lista de variables} BY {lista de variables}...../.....

Tipo de Variables : Alfanumérica o Numérica Discreta.

Tipo de Lista : Acepta especificaciones VARA TO VARX y admite hasta 200 variables (VARA TO VARX cuenta como 1)

Número máximo de categorías por Variable: 250

Número de listas de variables a tabular : 20 (todas aquellas que terminan por slash además de la última)

Número de celdas a generar para el total de tablas : $MAXCEL = \frac{(ESPACIO/4)}{D+2}$

D : dimensión de la tabla más grande

Dimensiones Máximas de una tabla : 10 (8 variables de control).

C.6. FASTABS : VARIABLES = {lista de variables} (mínimo, máximo)/..../
TABLES = {lista de variables} BY {lista de variables} ...../

Tipo de Variable : Numérica Discreta (entera)

Tipo de Lista : Primera parte (VARIABLES =) acepta especificación VARA TO VARX referida a la tarjeta VARIABLE LIST y admite hasta 100 variables nombradas implícita o explícitamente

Segunda parte (TABLES =) acepta especificación VARA TO VARX referida a la primera parte de la tarjeta (variables nombradas en ella) y admite hasta 100 variables (VARA TO VARX cuenta como 2)

Número máximo de Categorías por variable : 250

Número máximo de listas de tablas : 20 (aquellas que en la segunda parte terminan por slash además de la última)

Número máximo de Celdas a generar : MAXCEL= (ESPACIO/4)-(10xNVAR)
NVAR: es el número de variables nombradas explícita o implícitamente en la primera parte de la tarjeta FASTABS

Dimensión máxima de una tabla : 8 (6 variables de Control)

C.7. SCATTERGRAH {lista de variables} [(mínimo,máximo) / (LOWEST,máximo) / (mínimo,HIGHEST)] ./ {lista de variables} ...../{lista de variables} .....

Tipo de variable : Numérica Continua

Tipo de Lista : acepta especificación VARA TO VARX (ver punto 8.3.1.) y admite hasta 100 variables (mencionadas explícita o implícitamente)

Número máximo de
coeficientes a calcular : i) Si no se especifican opciones 1 ó 2 MCOEF=ESPACIO/24

ii) Si se especifican opciones 1 ó 2 MCOEF=7750

C.9. NONPAR CORR {lista de variables} WITH {lista de variables} /.....
/ lista de variables

Tipo de Variable : Numérica discreta

Tipo de Lista : Acepta especificaciones VARA TO VARX y admite hasta 100 variables nombradas explícita o implícitamente

Número máximo de listas de especificación de variables : 25

Número máximo de casos a procesar : Está dado por:

$$MXCAS = \frac{(ESPACIO/2)}{2xNV+1}$$

Si no se especifica la opción 1 ó 2 se debe agregar 4 unidades al denominador

NV: Si se usa la opción 2, es el número total de variables nombradas.
Si no se usa la opción 2, es el número de variables distintas. (Es decir si una variable aparece más de una vez se cuentan más de una vez, sólo si se usa la opción 2).

Número máximo de
Casos a procesar : Está dado por:

$$MXCAS = \frac{(ESPACIO/4) - 2 \times NV}{1 + NV}$$

NV: Número de variables nombradas explícita o implícitamente en la tarjeta SCATTERGRAM

Si se pasa de casos, no se cancela sino que ocupa los casos leídos hasta el máximo.

C.8. PEARSON CORR {Lista de variables} WITH {lista de variables) / {lista de variables}

Tipo de Variable : Numérica continua
Tipo de lista : Tiene dos tipos de listas (ver punto 10.1.1.) ambas aceptan especificaciones VARA TO VARX y admite hasta 250 elementos (esto incluye nombres de variables, las palabras TO y WITH y los de limitadores especiales, paréntesis derecho e izquierdo y los slash)

Limitación de
Variables : Un máximo de 500 variables pueden ser mencionadas en la tarjeta PEARSON CORR. Esta limitación no considera sólo las variables distintas sino que cuenta todos los nombres mencionados implícita o explícitamente

Número máximo de
listas de coeficiente : 40 (son aquellas que empiezan o terminan por slash)

E. TARJETAS DE CONTROL DEL O.S.

Las tarjetas de control del O.S. que son necesarias para realizar procesos de grabación están dadas esquemáticamente por la siguiente tabla:

LECTURA DE ARCHIVOS	TARJETA	DESCRIPCION EN:
Del Usuario	FT08	E.1.
Cintas		E.1.1.
Discos		E.1.2.
Del Sistema	FT03	E.2.
Cintas		E.2.1.
Discos		E.2.2.
GRABACION DE ARCHIVOS		
Del Sistema	FT04	E.3.
Cintas		E.3.1.
Discos		E.3.2.
Del Usuario	FT09	E.4.
Tarjetas		E.4.1.
Cintas		E.4.2.
Discos		E.4.3.

E.1. ectura de archivos del uario

E.1.1. Archivos en Cinta Magnética

La tarjeta de control es:

```
I              16
//FT08F001 DD  DSN=bbbb,UNIT=2400,LABEL=(n,SL),DISP=OLD,
//            DCB=(RECFM=FB,LRECL=XX,BLKSIZE=YYY),VOL=SER=ZZZ
```

Los argumentos del parámetro LABEL tienen el siguiente significado:

SL : indica que se usa LABEL STANDARD

n : indica la posición del archivo en el dispositivo magnético

Los bbbb que siguen al parámetro DSN deben ser reemplazados por el nombre del archivo.

Los XX que siguen al parámetro LRECL deben ser reemplazados por el largo del registro lógico.

Los YYYY que siguen al parámetro BLKSIZE deben ser reemplazados por el largo del bloque.

Los ZZZ que siguen al parámetro VOL=SER= deben ser reemplazado por el nombre del volumen.

E.1.2. Archivos en Discos

La tarjeta de control es:

```
1                16
//FT03F001 DD  DSN=bbb,UNIT=SYSDA,DISP=OLD,
//                VOL=SER=ZZZ
```

Los bbb y ZZZ que siguen a los parámetros DSN= y VOL=SER= respectivamente, tienen el mismo significado dado en el punto E.1.1.

E.2. Lectura de Archivos del Sistema

E.2.1. Archivos en Cinta

La tarjeta de control es:

```
1                16
//FT03F001 DD  DSN=SYSIN,UNIT=2400,LABEL=(n,SL),DISP=OLD,
//                DCB=BLKSIZE=4000,VOL=SER=ZZZ
```

Donde los ZZZ tienen el mismo significado dado en el punto E.1.1.

El n que aparece en el parámetro LABEL=, indica la posición física del archivo en la cinta, es decir si corresponde al primero, segundo, etc., archivo que graba.

E.2.2. Archivos en Discos

La tarjeta de control es:

```
1                16
//FT03F001 DD  DSNAME=bbb,UNIT=SYSDA,DISP=OLD,
//             VOL=SER=ZZZZ
```

Los bbb y los ZZZ que siguen a los parámetros DSNAME= y VOL=SER= tienen el mismo significado dado en el punto E.1.1.

E.3. Grabación de Archivos del Sistema

E.3.1. Grabación en Cinta

La tarjeta de control es:

```
1                16
//FT04F001 DD  DSN=SYSOUT,UNIT=2400,LABEL=(n,SL),DISP=(NEW,KEEP),
//             DCB=BLKSIZE=4000,VOL=SER=ZZZZ
```

Los ZZZ que siguen al parámetro VOL=SER= tienen el mismo significado dado en el punto E.1.1. y el n que sigue al parámetro LABEL es dado en el punto E.2.1.

E.3.2. Grabación en Discos

La tarjeta de control es:

```
1                16
//FT04F001 DD  DSN=bbbb,UNIT=SYSDA,DISP=(NEW,KEEP),
//             DCB=BLKSIZE=3510,VOL=SER=ZZZ,
//             SPACE=(3520,(YY,XX),RLSE)
```

Los bbb y los ZZZ que siguen a los respectivos parámetros tienen el mismo significado dado en el punto E.1.1.

El parámetro SPACE tiene como función reservar y crear espacio adecuado para almacenar el archivo. El largo de los bloques es standard e igual a 3520. Los YY deben ser reemplazados por el número de bloques que serán requeridos para almacenar el archivo. El usuario puede determinar este número mediante la fórmula siguiente:

$$YY = \frac{NVARS \times NCASOS}{750} + \frac{(NVARS \times 7) + (NSUBARCHIVOS \times 3) + (NROTVAR \times 13}{800} + \frac{NROTVAL \times 6}{800}$$

donde:

NVARS	: es el número de variables
NCASOS	: es el número de casos
NSUBARCHIVOS	: id. subarchivos
NROTVAR	: id. rótulos de variables
NROTVAL	: rótulos de valores de variables

Los ZZ deben ser reemplazados por un número igual al 25% de YY

E.4. Grabación de Archivos del Usuario

E.4.1. Archivos a Tarjetas

La tarjeta de control es:

```
1                16
//FT09F001 DD  SYSOUT=B,DCB=(RECFM=F,LRECL=80,BLKSIZE=80)
```

E.4.2. Archivos a Cintas

La tarjeta de control es:

```
1                16
//FT09F001 DD  DSN=bbb,UNIT=2400,LABEL=(n,SL),DISP=(NEW,KEEP),
//             DCB=(RECFM=FB,LRECL=80,BLKSIZE=800),VOL=SER=ZZZ
```

Los bbb y los ZZZ que siguen a los respectivos parámetros tienen el mismo significado dado en el punto E.1.1. y el n que sigue al parámetro LABEL el dado en el punto E.2.1.

E.4.3. Archivos a Discos

La tarjeta de control es:

```
1                16
//FT09F001 DD  DSN=bbb,UNIT=SYSDA,DISP=(NEW,KEEP),
//             DCB=(RECFM=FB,LRECL=80,BLKSIZE=2000),VOL=SER=ZZZ
```

Los bbb y los ZZZ que siguen a los respectivos parámetros tienen el mismo significado dado en el punto E.1.1.