UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

EL USO DE PAQUETES ESTADISTICOS PARA MODELOS LOGLINEALES

T E S I S

Que para obtener el título de:

A C T U A R I O

P r e s e n t a

JACQUELINE ENRIQUEZ BOLAÑOS

México, D. F.

1989

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

INTRODUCCION

Los métodos estadísticos se aplican generalmente a diversos campos tales como la biología, las ciencias sociales y la medicina para el análisis de datos que sean de interés para el investigador .

Uno de estos métodos para frecuencias de datos es el análisis mediante el cruce de clasificaciones o Tablas de Contingencia que se utilizan para para resumir resultados que generalmente provienen de experimentos, estudios clínicos, supervivencia, etc., inclusive, existen algunas publicaciones científicas del área de biología, que se han enfocado en el análisis de datos usando Tablas de Contingencia.

Estas tablas son apropiadas cuando las variables están medidas con una escala nominal u ordinal (categóricos).

El objeto de este trabajo es presentar las herramientas de cómputo que se pueden utilizar para este análisis, ya que para una gran cantidad de datos, resulta tedioso efectuar las operaciones y seguir los algoritmos manualmente para la obtención de resultados, por otra parte, si se toma en cuenta que a veces un análisis sugiere otro y por consiguiente el número de calculos a efectuar es mayor, el uso de la computadora agiliza los cálculos de una manera considerable, facilitando a los investigadores la tarea de

un análisis sustancial.

Para dar una visión sobre los aspectos más importantes sobre Tablas de Contingencia este trabajo se divide en dos partes, una parte correspondiente a la teoría y la otra al uso de los manuales de algunos paquetes de cómputo para poder efectuar el análisis de Tablas de Contingencia. La primera parte se divide en tres capítulos :

El primer capítulo explica qué es una Tabla de Contingencia y como se forma.

El segundo capítulo se basa fundamentalmente en la teoría expuesta por Fienberg (1977) en el libro "The Analysis of Cross-Classified Categorical Data".

El tercer capítulo está basado en general en los artículos de Nelder y Wedderburn (1972) y Nelder (1974) en donde los modelos loglineales se ven como un caso particular de los modelos lineales clásicos.

Pasando a un aspecto más práctico, la segunda parte en sus tres capítulos corresponden a 3 paquetes estadísticos : SYSTAT, BMDP y GLIM que son los más usuales y comerciales, así como su uso y las ventajas y desventajas que cada uno ofrece.

PRIMERA PARTE

1. TABLAS DE CONTINGENCIA

Supóngase que se tiene una población con n individuos y que cada uno es descrito por un número de atributos. Todos los individuos con la misma descripción son contados y clasificados de acuerdo a los criterios de interés para el investigador, este valor pertenece a una celda de la tabla formando así una Tabla de Contingencia, construida generalmente con el fin de establecer las relaciones entre las variables involucradas.

Cuando las celdas son definidas en términos de categorías de dos o más variables se determina una estructura. La estructura "natural" para dos variables, p.e., es a menudo un arreglo rectangular con columnas correspondientes a las categorías de una variable y renglones para las categorías de la segunda variable. La posición de las celdas refleja de alguna manera las características de los individuos que pertenecen a ellos. Por

ejemplo, los individuos de una celda específica tienen una característica común con individuos de todas las celdas del mismo renglón y otra característica en común con todos los individuos en todas las celdas de la misma columna. Un buen modelo matemático deberá reflejar esta estructura.

Supóngase que los elementos de una población pueden ser clasificados de acuerdo a dos criterios o variables. Esto es p.e. si una población está formada por alumnos de diferentes escuelas que aspiran ingresar a la UNAM , el criterio de clasificación podría ser en primer lugar la escuela de donde provienen, este criterio tiene varias categorías : CCH, PREPARATORIA, VOCACIONAL, OTRA INSTITUCION. Un segundo criterio de clasificación podría ser la calificación que obtienen en el examen para ingresar a la UNAM, este criterio tiene también varias categorías que se pueden agrupar de acuerdo al interés del investigador, una de las clasificaciones podría ser p.e. : 6, 6.0 - 7.9, 8.0 - 10.

La clasificación dada por estos criterios (o variables) debe ser exhaustiva, esto es, la clasificación elegida agrupa en sus diferentes categorías a todos los miembros de la población; y mutuamente exclusivas, es decir no puede quedar clasificado en 2 categorías a la vez de una variable: un miembro de la población pertenece a una sola celda.

Esta clasificación (exhaustiva y mutuamente exclusiva) es un requisito que deben cumplir las variables para el análisis en Tablas de Contingencia.

La pregunta natural que surge en el ejemplo anterior es: la calificación que obtienen los aspirantes a la UNAM es independiente de la escuela que provienen ? i.e. influye un criterio en el otro ? , y aún mucho más, en caso de no haber independencia como se relacionan o asocian un criterio y el otro?.

Para resolver este tipo de preguntas y otras más se utiliza el análisis de Tablas de Contingencia.

Para el análisis de Tablas de Contingencia, en la actualidad son ampliamente usados los modelos loglineales.

Las Tablas de Contingencia más simples son las de 2x2 (dos criterios con dos categorías cada una), pudiéndose extender a r-criterios con las categorías necesarias para cada uno.

En la tabla 1-1 se presenta una Tabla de Contingencia de 4x2, en donde las entradas son frecuencias o conteos correspondientes a los datos obtenidos al realizar un cuestionario con alumnos de una escuela a nivel bachillerato de acuerdo al turno en el que estaban inscritos y a si habían recibido información por su cuenta para seleccionar las materias que cursarían el siguiente semestre.

Las clasificaciones indicadas pertenecen a una muestra de 2115 alumnos con 2 variables: si ha recibido información (si o no) y turno (01, 02, 03, 04).

tabla 1-1

información / turno	01	02	03	04
si	183	134	194	113
no	305	391	404	391

total 2115 alumnos

Para el análisis de Tablas de Contingencia, como se señaló anteriormente, es común usar los modelos logarítmicos lineales, que son modelos lineales en los logaritmos de las frecuencias esperadas en las celdas. Estos modelos son similares a los utilizados en el análisis de varianza. Por eso es que algunos autores han adoptado para Tablas de Contingencia el término de "interacción" refiriéndose a la asociación entre las variables, i.e. las relaciones estructurales entre las variables categóricas, por consiguiente se hablará de interacciones de 1o. orden entre pares de variables, interacciones de 2o. orden entre tres variables y así sucesivamente. Este tipo de asociaciones se presentan cuando la hipótesis de independencia no es aceptada, y por consiguiente el modelo de completa independencia no se ajusta bien a los datos.

Estos modelos se mostrarán en las siguientes secciones, empezando por los más sencillos, para dos y tres criterios y posteriormente para n-criterios.

2. MODELOS LOGARITMICOS LINEALES

2.1 Tablas de dos criterios

El interés en esta sección son las Tablas de Contingencia de 2 criterios, esto es, explorar las relaciones entre dos variables o criterios (correspondientes a renglones y columnas de la tabla).

La tabla más simple de dos criterios es la de 2 variables con 2 categorías. La generalización a i-categorías por j-categorías puede ser vista de la siguiente forma:

		Variable 2					
		1	2	3 , ... ,	J	total	
	1	x_{11}	x_{12}	....	x_{1j}	$x_{1.}$	
V	2	x_{21}			x_{2j}	$x_{2.}$	
a	3	x_{31}			x_{3j}	$x_{3.}$	
r							
i							
a	.						
b	.						
l	.						
e							
i							
	I	x_{i1}	x_{i2}	x_{i3} ...	x_{ij}	$x_{i.}$	
total		$x_{.1}$	$x_{.2}$	$x_{.3}$ ...	$x_{.j}$	$x_{..}$ = N	

Donde

$$x_{i.} = \sum_{j=1}^{J} x_{ij} \qquad (2.1.1)$$

$$x_{.j} = \sum_{i=1}^{I} x_{ij} \qquad (2.1.2)$$

$$x_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij} = \sum_{i=1}^{I} x_{i.} = \sum_{j=1}^{J} x_{.j} \qquad (2.1.3)$$

$x_{i.}$
$\}$ representan los totales marginales de las categorías i y j
$x_{.j}$

$x_{..}$ representa el número total de observaciones

Los valores esperados de x_{ij} son $m_{ij} = NP_{ij}$, donde P_{ij} es la probabilidad de que una observación pertenezca a la celda (i,j). Bajo la hipótesis de independencia de renglones y columnas se tiene:

$$P_{ij} = P_{i.} P_{.j} \qquad (2.1.4)$$

por lo tanto, bajo la hipótesis de independencia

$$m_{ij} = NP_{i.} P_{.j} \qquad (2.1.5)$$

esto es

$$m_{ij} = \frac{N\, m_{i.} m_{.j}}{N^2}$$

de donde

$$\hat{m}_{ij} = \frac{x_{i.} x_{.j}}{N} \qquad (2.1.6)$$

Esta hipótesis de independencia se puede probar con la estadística X^2 de Pearson :

$$X^2 = \sum \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}} \qquad (2.1.7)$$

o con

$$G^2 = 2 \sum (\text{observados}) \log \left(\frac{\text{observados}}{\text{esperados}}\right)$$

En ambos casos la suma es sobre toda la tabla.

La tabla 2.1-I es un ejemplo de una Tabla de Contingencia de 3x3 categorías, está basada en una encuesta a un grupo de médicos realizada en una ciudad. Se les dividió en 3 categorías en base a su currículum académico y sus ingresos al cabo de 15 años de ejercicio para analizar si existe alguna relación entre las dos variables. La tabla 2.1-Ia muestra los valores esperados bajo la hipótesis de independencia (2.1.4). Al calcular la estadística X^2 se obtiene que es igual a 6.11, comparando este valor con el obtenido en tablas, se tiene que se acepta la hipótesis de independencia, esto es que no existe alguna relación entre el currículum académico y los ingresos obtenidos.

Regresando a (2.1.6) y habiendo mostrado como se obtienen los valores estimados para 2 criterios bajo la hipótesis de independencia, ahora se desea obtener un modelo que incluya los términos involucrados. Para Tablas de Contingencia se utilizan los modelos loglineales , que postulan que los valores esperados de las observaciones están dadas por una combinación lineal de ciertos parámetros.

Tabla 2.1-1

CURRICULUM	INGRESO alto	mediano	bajo	
alto	18	17	5	$x_{1.}=40$
mediano	26	38	16	$x_{2.}=80$
bajo	6	15	9	$x_{3.}=30$
	$x_{.1}=50$	$x_{.2}=70$	$x_{.3}=30$	$x_{..}=N=150$

(a) Valores esperados

CURRICULUM	INGRESO alto	mediano	bajo	
alto	13.3	18.7	8.0	$m_{1.}=40$
mediano	26.7	37.3	16.0	$m_{2.}=80$
bajo	10.0	14.0	6.0	$m_{3.}=30$
	$m_{.1}=50$	$m_{.2}=70$	$m_{.3}=30$	$m_{..}=150$

El Modelo loglineal

Tomando logaritmos de (2.1.6), bajo la hipótesis de independencia se tiene:

$$\log \hat{m}_{ij} = \log x_{i.} + \log x_{.j} - \log N \qquad (2.1.8)$$

Esta forma puede tener una similitud a la notación del análisis de varianza, aunque en los modelos loglineales los términos involucrados en el modelo son empleados para describir las relaciones estructurales entre las variables categóricas correspondientes a las dimensiones de la tabla.

De esta manera (2.1.8) se puede ver como:

$$\log m_{ij} = u + u_1(i) + u_2(j) \qquad (2.1.9)$$

que es el modelo de completa independencia para tablas de dos criterios, donde :

$$u = \frac{1}{IJ}\sum_{i}^{I}\sum_{j}^{J} \log m_{ij} \qquad (2.1.10)$$

Representa la media general de los logaritmos de los valores esperados.

$$u_1(i) = \frac{1}{J}\sum_{j}^{J} \log m_{ij} - u \qquad (2.1.11)$$

$$u_2(j) = \frac{1}{I}\sum_{i}^{I} \log m_{ij} - u \qquad (2.1.12)$$

Representan desviaciones de la media general.

A menudo sucede que el modelo de independencia no ajuste bien a los datos así que se añade el término de interacción o asociación $u_{12}(ij)$ entre la variable 1 y la variable 2 al modelo de completa independencia 2.1.9

$$\log m_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij) \qquad (2.1.13)$$

Entonces $u_{12}(ij)$ representa los efectos de interacción entre los niveles i y j de las variables 1 y 2 respectivamente.

Si $u_{12}(ij) = 0$ el modelo referido es el (2.1.9)

De aquí en adelante, para simplificar la notación, se

entenderá por ejemplo que

$u_{iz}(ij)$ simplemente como u_{iz}

quedando claro que los subindices i y j (en este caso) de las variables están implícitas al referirse a u_{iz} .

2.2 Modelos de muestreo

Generalmente los datos de las frecuencias de las Tablas de Contingencia (de cualquier dimensión) provienen de alguno de los siguientes tres modelos de muestreo :

POISSON : Se observa un conjunto de procesos Poisson uno para cada celda sobre un período de tiempo sin un conocimiento a priori respecto al número de observaciones i.e. no se fija el tamaño de la muestra.

La función de densidad de probabilidad es:

$$f([x]) = \prod \frac{m^{x} e^{-m}}{x!}$$

MULTINOMIAL : Se toma una muestra de tamaño N y se clasifica cada miembro de la muestra de acuerdo a sus valores según las variables que se tienen. La

función de densidad de probabilidad es:

$$f(\{x\}) = \frac{N!}{\prod x!} \prod \left(\frac{m}{N}\right)^{x}$$

MULTINOMIAL PRODUCTO : Aunque en los estudios de las observaciones solamente una muestra puede ser examinada, en situaciones experimentales es usual tener varios grupos. Cuando miembros de diferentes grupos no están apareados, las frecuencias de las celdas representan individuos y los totales de grupos forman los totales marginales.

La función de densidad depende de la configuración que se esté tratando, p.e. si se está estudiando la configuración $x.jk$ la distribución marginal correspondiente sería:

$$f(\{x.jk\}) = \frac{N!}{\prod_{jk} x.jk!} \prod_{jk} \left(\frac{m.jk}{N}\right)^{x.jk}$$

Birch (1963), Haberman (1974) y otros autores han demostrado que los estimadores de máxima verosimilitud de los valores esperados de las celdas bajo los modelos loglineales que se consideren, son los mismos bajo los tres esquemas señalados anteriormente. La única condición requerida para este resultado es que los términos "u" correspondientes a los totales marginales

fijos en el esquema Multinomial-Producto sean incluídos en el modelo loglineal bajo consideración.

2.3 Tablas de tres criterios

Tablas de I x J x K

A medida que el número de criterios que describe a cada individuo aumenta, el análisis e interpretación resulta más complicado. En esta sección se describirán brevemente las Tablas de Contingencia para tres criterios.

Como un ejemplo ilustrativo de una Tabla de Contingencia de tres criterios, considérense los datos de la tabla 2.3-1. En ella se tiene la población del año 1989 de estudiantes del CCH que trabajan como obrero o patrón, clasificado por: plantel en el que estudian (PLANTEL): Azcapotzalco, Naucalpan, Oriente, Sur, Vallejo; tipo de ingreso (INGRESO): si es de primer ingreso o de reingreso; y ocupación (OCUP) considerando sólo 2 ocupaciones en este estudio, en el sentido de que si trabaja como empresario , empleador, patrón o como obrero, empleado. Los nombres que están entre paréntesis y con mayúsculas, son los que se utilizarán más adelante para hacer referencia a ellas.

Estas clasificaciones se tomaron en base a un estudio para determinar cuales son las relaciones que existen entre estas variables. De esta manera surgirían varias preguntas, como por

ejemplo: El plantel en el que estudia un alumno que trabaja, tiene alguna asociación con el tipo de ocupación? El tipo de ocupación, la forma de ingreso al plantel y el plantel en el que estudia el alumno que trabaja, son completamente independientes?

Cuando se tienen 3 criterios o variables de clasificación como en el ejemplo anterior, se forma una Tabla de Contingencia de tres dimensiones.

El análisis de estas tablas presentan nuevos problemas conceptuales comparado con el de tablas bi-dimensionales. Este se expondrá de manera general en esta sección.

Siguiendo la notación de Tablas bi-dimensionales (2.1.1 - 2.1.3), donde el número de índices corresponde al número de variables, se extiende a tablas de 3 dimensiones. Para éstas,

Tabla 2.3-1

		AÑO 1980 INGRESO			
		primer ingreso		reingreso	
PLANTEL	OCUP	patrón	obrero	patrón	obrero
Azcapotzalco		18	822	37	1567
Naucalpan		22	622	30	1190
Oriente		21	855	38	1807
Sur		21	663	38	1412
Vallejo		33	901	60	2312

los totales marginales son :

$$x_{i..} = \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk}$$

$$x_{.j.} = \sum_{i=1}^{I} \sum_{k=1}^{K} x_{ijk}$$

$$x_{..k} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ijk}$$

$$x_{...} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk}$$

La probabilidad de que una observación pertenezca a la celda (i,j,k) es P_{ijk} y los valores observados son m_{ijk}, donde $i = 1,2,\ldots,I$, $j = 1,2,\ldots,J$, $k = 1,2,\ldots,K$. Bajo la hipótesis de completa independencia[1] se tiene que (análogo a 2.1.4 -2.1.6):

$$P_{ijk} = P_{i..} P_{.j.} P_{..k} \qquad (2.3.1)$$

$$\hat{m}_{ijk} = \frac{x_{i..} x_{.j.} x_{..k}}{N^2} \qquad (2.3.2)$$

(2.3.2) es el estimador de los valores esperados cuando se supone independencia entre las tres variables. Tomando logaritmos se tiene :

$$\log \hat{m}_{ijk} = \log x_{i..} + \log x_{.j.} + \log x_{..k} - 2N$$

1

Al hablar de completa independencia quiere decir que:
Sean a,b,c tres eventos y p la probabilidad de que ocurran, entonces:
$p(abc)=p(a)p(b)p(c)$; $p(ab)=p(a)p(b)$; $p(ac) = p(a)p(c)$
$p(bc) = p(b)p(c)$.

Partiendo de este modelo se puede ver el modelo loglineal para tres criterios.

Modelos loglineales

Utilizando la notación del análisis de varianza, para log mijk, el modelo loglineal de completa independencia es :

$$\log m_{ijk} = u + u_1 + u_2 + u_3 \qquad (2.3.3)$$

donde :

$$u = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log m_{ijk} \qquad (2.3.4)$$

$$u_1 = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \log m_{ijk} - u \qquad (2.3.5)$$

$$u_2 = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \log m_{ijk} - u \qquad (2.3.6)$$

$$u_3 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log m_{ijk} - u \qquad (2.3.7)$$

Como se puede ver:

(2.3.4) Es la media general de los logaritmos de los valores esperados.

(2.3.5) (2.3.6) (2.3.7) Son las desviaciones de la media general.

Pero, qué pasa si las variables que se están estudiando no son independientes entre sí?. Frecuentemente el modelo de completa

independencia no es el que mejor se ajusta a los datos ya que éste no contempla las posibles asociaciones entre las variables, pudiendo encontrarse en alguno de los siguientes cuatro casos :

1. Independencia de una variable con las otras dos conjuntamente. En este caso se tienen los siguientes modelos:

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{12}$

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{13}$

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{23}$

2. Independencia condicional de dos variables dada la otra. Los modelos correspondientes a este caso son:

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{12} + u_{13}$ (2.3.8)

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{12} + u_{23}$

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{13} + u_{23}$

3. Asociación entre los tres pares de variables sin que la relación esté afectada por la otra variable. Teniendo el modelo

   $\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{12} + u_{13} + u_{23}$ (2.3.9)

4. Asociación entre las tres variables. El modelo en este caso es

$$\log m_{ijk} = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123} \qquad (2.3.10)$$

Como se puede observar los casos 1,2 y 3 corresponden a modelos particulares de (2.3.10). Por ejemplo el modelo (2.3.8) se obtiene del modelo (2.3.10) al considerar las siguientes hipótesis:

$$H_0 = u_{23} = 0, \quad u_{123}=0$$

Al modelo (2.3.10) se le suele llamar modelo saturado, ya que contiene un parámetro para cada celda, es decir, el número de parámetros en el modelo es igual al número de celdas. En el caso de rxc el modelo saturado es el (2.1.13).

Los modelos que se van a consideran son los llamados modelos jerárquicos. Esto es, siempre que un término se incluya, los términos con efectos de orden menor involucrados deberán estar también incluídos.

Por ejemplo, si un modelo incluye el término u_{12} los términos u_1 y u_2 deben estar incluídos en el modelo. En el caso de incluir el término u_{123}, los términos u_{12}, u_{13}, u_{23} también deberán aparecer en el modelo.

Si se consideran los modelos

$$\log m_{ijk} = u + u_1 + u_2 + u_{123} \qquad (2.3.11)$$

$$\log m_{ijk} = u + u_1 + u_2 + u_3 + u_{123} \qquad (2.3.12)$$

$$\log m_{ijk} = u + u_1 + u_2 + u_3 + u_{13} \qquad (2.3.13)$$

Se puede ver que los modelos (2.3.11) y (2.3.12) no cumplen con el principio de jerarquía solamente el (2.3.13) es un modelo jerárquico.

Valores esperados.

Considerando el modelo de completa independencia (2.3.3), se señaló como se obtuvieron en forma directa los valores esperados

$$\hat{m}_{ijk} = \frac{x_{i..} x_{.j.} x_{..k}}{N^2}$$

que es el estimador de máxima verosimilitud para el modelo (2.3.3).

Si se tiene el modelo

$$\log m_{ijk} = u + u_1 + u_2 + u_3 + u_{13} + u_{23} \qquad (2.3.14)$$

esto es

$$u_{12} = u_{123} = 0$$

se tendría que los valores esperados estimados para este modelo son de la forma

$$\hat{m}_{ijk} = \frac{x_{i.k} \; x_{.jk}}{x_{..k}} \qquad (2.3.15)$$

obtenidos directamente ya que

$$m_{.jk} = e^{u + u_2 + u_3 + u_{23}} \sum e^{u_1 + u_{13}} \qquad (2.3.16)$$

$$m_{i.k} = e^{u + u_{i} + u_{3} + u_{i3}} \Big/ \sum e^{u_{2} + u_{23}} \qquad (2.3.17)$$

$$m_{..k} = e^{u + u_{3}} \Big/ \sum e^{u_{i} + u_{2} + u_{33} + u_{23}} \qquad (2.3.18)$$

dividiendo el producto de (2.3.16) y (2.3.17) entre (2.3.18) se tiene

$$m_{ijk} = \frac{m_{i.k}\ m_{.jk}}{m_{..k}}$$

esto es, para cada valor ajustado de k se tiene independencia para cada tabla marginal IxJ correspondiente a este valor ajustado.

De la misma manera, para el modelo

$$\log m_{ijk} = u + u_{i} + u_{2} + u_{3} + u_{i3} \qquad (2.3.19)$$

se tendría que

$$\hat{m}_{ijk} = \frac{x_{.j.}\ x_{i.k}}{N} \qquad (2.3.20)$$

La tabla 2.3-2 presenta los valores esperados obtenidos con (2.3.20) tomando en cuenta los datos de la tabla 2.3-1 bajo el modelo (2.3.19), de esta manera $\hat{X}^2 = 9.09$.

Desafortunadamente no todos los estimadores se pueden obtener en forma directa, como en el caso del modelo (2.3.9), y es por esta razón que hay que recurrir a métodos numéricos para el cálculo de los valores esperados.

Tabla 2.3-2

		INGRESO			
		primer ingreso		reingreso	
PLANTEL	OCUP	patrón	obrero	patrón	obrero
Azcapotzalco		21.36	816.64	40.89	1563.11
Naucalpan		16.57	633.43	31.10	1188.90
Oriente		22.33	853.67	47.08	1799.92
Sur		17.44	666.56	36.96	1413.04
Vallejo		23.81	910.19	60.46	2311.54

Estos métodos son diferentes, y ya que no es el propósito de esta Tesis el describirlos, solamente se hará referencia que para los paquetes BMDP y SYSTAT (especificados en la II Parte de esta Tesis) usan el Método Iterativo de Ajuste Proporcional[2], y el paquete GLIM (también especificado en la II Parte) utiliza el método de Newton-Raphson[3]. Es importante denotar que en ambos casos se llegan a los mismos valores esperados.

Hasta ahora se ha tratado únicamente con Tablas de Contingencia de dos y tres criterios de clasificación, ahora se presentará la generalización a n criterios de clasificación.

[2] Para mayores detalles de este método ver: Fienberg, "The Analysis of Cross-Clasified Categorical Data", p 33-36.

[3] Para mayores detalles de este método ver: Agresti, "Analysis of Ordinal Categorical Data", p 237-241.

2.4 Modelos para cuatro o más criterios

De manera similar, como se mostraron los modelos para Tablas de Contingencia de 2 y 3 criterios se puede generalizar para n-criterios. Los valores observados en celdas, por ejemplo, de cuatro criterios tendrían cuatro subíndices i,j,k,l.

Con 2 criterios el modelo loglineal saturado (2.1.13) consta de 4 términos , con tres criterios de 8 $(u, u_1, u_2, u_3, u_{12}, u_{13}, u_{23}, u_{123})$ y en general para n dimensiones el modelo loglineal saturado tendrá 2^n términos "u" : n u-términos de los efectos principales $(u_1, u_2, \ldots, u_n)$ y otro término que representa la media general; más todas las combinaciones posibles de 2 variables: C_2^n que forman los términos de dos factores y así sucesivamente. De esta manera se tiene que en general para obtener el número de términos de r-factores será de C_r^n .

Para 2 dimensiones se tienen 4 modelos jerárquicos que se pueden tomar en cuenta para efectuar el ajuste, este número crece muy rápido al aumentar la dimensión. En el caso de 4 dimensiones, se tienen 113 modelos jerárquicos, todos incluyendo los u-términos de los efectos principales: $u_1, u_2, \ldots$. Good [1975] se dió a la tarea de enumerar todos los posibles modelos de independencia (ambos mutuamente y condicionalmente independientes) en una tabla n-dimensional. El número de modelos crece rápidamente : en una tabla de 10 dimensiones hay 3 475 978 modelos (este número es pequeño comparado con el no. de modelos jerárquicos para 10 dimensiones).

También es posible encontrar estimadores directos para determinados modelos en el caso de 4 o más criterios. Por ejemplo, para tablas de cuatro dimensiones, entre los modelos donde los estimadores se pueden obtener directamente están los siguientes:

$$\log m_{ijkl} = u + u_{1} + u_{2} + u_{3} + u_{4}$$

$$\log m_{ijkl} = u + u_{1} + u_{2} + u_{3} + u_{4} + u_{23} + u_{34} + u_{234} + u_{124}$$

los estimadores directos respectivamente son :

$$\hat{m}_{ijkl} = \frac{x_{i\ldots}\, x_{.j.}\, x_{..k}\, x_{...l}}{N^{3}}$$

$$\hat{m}_{ijkl} = \frac{x_{.kl}\, x_{.jkl}}{x_{..kl}}$$

Sin embargo para modelos que se deseen considerar y no existan estimadores directos se pueden emplear los métodos que ya se mencionaron.

2.5 Necesidad de utilizar modelos para tablas con más de dos criterios

Podría pensarse el por qué si para Tablas de Contingencia de 3 o más criterios el análisis de los datos y su interpretación presenta problemas especiales, por qué no usar tablas marginales de dos criterios y analizar las relaciones entre pares de

variables.

Hasta años recientes (principios de los 70's), las técnicas estadísticas y computacionales disponibles para el análisis de Tablas de Contingencia eran muy limitados y muchos investigadores manejaban varias tablas de 2 criterios utilizando los totales marginales de la tabla multidimensional. El análisis de esta manera (que sólo en algunos casos dan una buena visión) surgían los siguientes problemas:

1) Confusión de la relación marginal entre un par de variables con la relación entre ellas cuando otra está presente.
2) No permite el análisis simultáneo de la relación entre pares de variables.
3) Ignora la posibilidad de asociación entre más de dos variables.

Como ejemplo se presentan los datos de la tabla 2.5-1 la cual

Tabla 2.5-1

Lugar donde recibieron cuidado	Cuidado prenatal	Sobrevivencia de infantes muertos	sobrevivientes
clínica A	menos	3	176
	más	4	293
clínica B	menos	17	197
	más	2	23

fuente: Bishop [1969]

Tabla 2.5-2

Cuidado prenatal	Sobrevivencia de infantes: muertos	sobrevivientes
menos	20	373
más	6	316

contiene datos analizados por Bishop [1969], que relaciona la sobrevivencia de infantes de acuerdo al cuidado prenatal que recibieron por las madres, clasificado como 'menos' o 'más'. Las madres fueron atendidas en 2 clínicas: clínica A o clínica B.

Se desea conocer si existe asociación entre la sobrevivencia del infante y el cuidado que éste recibió. Analizando estos datos bajo el modelo

Ho: LUGAR + CUIDADO + SOBREVIVENCIA + CUIDADO.SOBREVIVENCIA

se obtiene que $X^2 = 184.00$ con 3 grados de libertad, comparando este valor con el obtenido en tablas se llega a la conclusión de que se rechaza el modelo Ho, esto es, la sobrevivencia no está relacionada con el cuidado del infante.

Por otra parte, si se analizan los totales marginales (tabla 2.5-2), bajo el modelo de independencia, se tiene que $X^2 = 5.612$ con 2 g.l, concluyendo erróneamente que la sobrevivencia está relacionada con el cuidado prenatal.

2.6 Estadísticas para bondad de ajuste

Ya que se obtuvieron los valores esperados estimados para el modelo loglineal escogido, se puede probar que tan buenos son estos valores usando alguna de las siguientes estadísticas:

$$X^2 = \sum \frac{(\text{observados} - \text{estimados})^2}{\text{estimados}} \qquad (2.6.1)$$

$$G^2 = 2 \sum (\text{observados}) \log \left(\frac{\text{observados}}{\text{estimados}} \right) \qquad (2.6.2)$$

Cuando el tamaño de la muestra es grande se tiene que X^2 y G^2 se distribuyen como una χ^2 con los siguientes grados de libertad:

grados de libertad = No.celdas -No.parámetros ajustados linealmente independientes

Por ejemplo, para una tabla de 3 dimensiones (IxJxK) para el modelo

$$\log m_{ijk} = u + u_1 + u_2 + u_3 + u_{12} \qquad (2.6.3)$$

el número de parámetros estimados ajustados es

	No. parámetros
u	1
u_1	(I-1)
u_2	(J-1)
u_3	(K-1)
u_{12}	(I-1)(J-1)

entonces los grados de libertad asociados al modelo (2.6.3) son

$$g.l. = (IJK) - [1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1)]$$
$$= [(K-1)(IJ-1)]$$

Las estadísticas X^2 y G^2 son equivalentes asintóticamente, esto es que son equivalentes en muestras grandes cuando la hipótesis nula es verdadera.

La G^2 es menos familiar que la X^2. La desventaja de la G^2 es que es más fácil el cálculo de la X^2 y la ventaja es que la G^2 es la estadística que es minimizada por los estimadores de máxima verosimilitud y por otra parte se puede particionar como se señala posteriormente.

2.7 Selección del modelo

Como se mencionó anteriormente, para una tabla de 3 criterios se tienen 8 posibles modelos jerárquicos y a medida que aumenta el número de criterios, el número de modelos se incrementa rapidamente siendo de mayor complejidad e involucrando mayor número de parámetros. El problema ahora es, qué modelo es el mejor? o qué modelo es el más adecuado?. A menudo es preferible un modelo sencillo que uno complicado que provea un mejor ajuste.

Las pruebas de bondad de ajuste permiten ver que tan cercanos están los valores esperados de los valores observados para un

modelo en particular. Para la selección del "mejor" modelo , no se pueden hacer pruebas de hipótesis de cada modelo independientemente y de esta manera efectuar la selección. En este caso se tiene que recurrir a métodos establecidos .

Bishop [1969], Brown [1976], Fienberg [1970], Goodman [1970, 1971] y Ku Kullback [1968] han seguido métodos con diferentes aproximaciones al problema de la selección del modelo.

El método que en forma general se describe en este trabajo y que se puede utilizar para escoger el "mejor" modelo que en la práctica refleje lo más adecuadamente posible las interacciones entre las variables, es el que consiste en particionar la estadística de bondad de ajuste G^2 para un modelo jerárquico, en varias partes aditivas y cada parte tiene una distribución asintótica χ^2 con sus correspondientes grados de libertad.

Estas particiones se obtienen considerando un conjunto de modelos loglineales jerárquicos "anidados", o sea que al ser ordenados los términos del primer modelo están incluídos en el segundo (de mayor complejidad que el primero) y los del segundo modelo en los del tercero y así sucesivamente, hasta llegar al que el investigador decida, dado que éste tiene información auxiliar y al menos tiene conocimiento, de alguna manera, que modelos pueden acercarse al que busca.

Un conjunto de estos modelos loglineales jerárquicos anidados para una tabla de 3 criterios puede ser:

a) $u + u_1 + u_2 + u_3$

b) $u + u_1 + u_2 + u_3 + u_{12}$

c) $u + u_1 + u_2 + u_3 + u_{12} + u_{13}$

d) $u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23}$

Si se cambia algún término y se sigue cumpliendo que sean anidados se generaría otro conjunto de modelos diferentes.

Un resultado de interés es el que señala que con la estadística

$$2 \sum (\text{observados}) \log \left[\frac{(\text{esperados})\text{mod1}}{(\text{esperados})\text{mod2}} \right] \qquad (2.7.1)$$

se puede probar si la diferencia entre los valores esperados de dos modelos (mod1 y mod2) se debe a una simple variación aleatoria dado que los valores esperados satisfacen el modelo 1.

Se denotará como G(a), G(b), G(c), G(d) a la estadística de verosimilitud para los modelos a, b, c, d respectivamente , Un resultado que se cumple en general es

$$G^2(a) >= G^2(b) >= G^2(c) >= G^2(d)$$

Este resultado es importante y es por una de las razones que se utiliza esta estadística en lugar de particionar la X^2, ya que este resultado en general no es cierto para cualquier conjunto anidado de modelos.

Además las estadísticas $G^2(a) - G^2(b)$, $G^2(b) - G^2(c)$, $G^2(c) - G^2(d)$ son de la forma señalada anteriormente en (2.7.1) y pueden ser usadas para probar diferencias entre los modelos a y b, b y c, c y d.

El método de selección en general consiste en tomar del conjunto de modelos jerárquicos anidados que se seleccionaron, las diferencias de las estadísticas G^2 del modelo menos simple con el que le sigue de más simplicidad. Esta diferencia se prueba con la estadística G^2 con los grados de libertad igual a diferencia de los grados de libertad de los dos modelos involucrados. Si esta prueba es no significativa se desecha el modelo menos simple de los 2 y se procede a repetir este procedimiento. Si la prueba es significativa entonces el modelo que se tomará es el menos simple entre los 2 que se tomó la diferencia y éste es el que se considerará como el "mejor" modelo que ajusta los datos.

3. MODELOS LINEALES GENERALIZADOS

3.1 Modelos Lineales Generalizados

Los Modelos loglineales señalados anteriormente son un caso particular de los Modelos Lineales Generalizados, que pueden caracterizarse por involucrar una variable de respuesta y un grupo de variables explicativas. Esta sección describirá de manera general, los aspectos básicos para el análisis de Tablas de Contingencia[4], y se basa fundamentalmente en los artículos de Nelder.

Nelder J.A. [1974] menciona que los Modelos Lineales Generalizados son una extensión de los Modelos Lineales Clásicos. Es por esto que a continuación se señalan las principales

[4]Para una explicación más amplia de Modelos Lineales Generalizados y referencias adicionales, ver Nelder y Wedderburn [1972].

características del Modelo Lineal Clásico y partiendo de esto se verá como se define el Modelo Lineal Generalizado.

El Modelo Lineal Clásico se podría definir mediante tres componentes:

1. EL COMPONENTE ALEATORIO O ESTRUCTURA DE ERROR.
2. EL COMPONENTE SISTEMATICO O PREDICTOR LINEAL.
3. LA FUNCION LIGA.

1. EL COMPONENTE ALEATORIO O ESTRUCTURA DEL ERROR

Se tiene un vector $Y = (y_1, y_2, \ldots, y_n)$ de variables aleatorias independientes normalmente distribuídas con medias $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ con $E(Y) = \mu$ donde $\mu = Xb$ esto es Y se distribuye como una Normal con varianza cte. σ^2 y $E(Y) = \mu$.

Una suposición que hay que añadir, es que en los Modelos Lineales Clásicos los errores ε_i son independientes y se distribuyen como una Normal con varianza constante σ^2.

2. EL COMPONENTE SISTEMATICO O PREDICTOR LINEAL

Establece la forma en que las variables explicativas (x's) se usan para "predecir" el valor de la respuesta, su forma general es:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\,\beta_j \qquad i=1,\ldots,n \qquad o \qquad \eta = X\beta$$

donde x_{ij} es el valor de la j-ésima variable para la observación i y las bj son parámetros a estimarse ya que generalmente son

desconocidos. La matriz X, del orden de nxp es llamada "matriz de diseño". Las variables explicativas $(x_1, x_2, \ldots, x_p)$ generan un predictor lineal dado por η.

3. LA FUNCION LIGA

Relaciona la media de la i-ésima observación y_i su componente sistemático mediante una función "g". Si se denota por μ_i a la media de Y_i (variable de respuesta) se tiene que

$$\eta_i = g(\mu_i)$$

Para los Modelos Lineales Clásicos el valor esperado de Y y el predictor lineal son iguales:

$$E(Y) = \mu$$

$$g(\mu) = X\beta = \mu \qquad \text{i.e. } \eta = \mu$$

Entonces la función liga es la identidad.

Partiendo de estos 3 conceptos de los Modelos Lineales Clásicos, los Modelos Lineales Generalizados permiten 2 extensiones:

1. Están definidos para miembros de la familia exponencial, e.e. Normal, Poisson, Binomial, Gamma.

   Como se señaló en la sección 2.2, los modelos de muestreo para modelos loglineales son las distribuciones Poisson, Multinomial, Multinomial Producto y cualquiera de éstas llevan a los mismos estimadores. Por lo consiguiente para Tablas de Contingencia en los Modelos Lineales Generalizados se tomará

la distribución Poisson.

2. La función liga puede ser cualquier función monótona diferenciable.

Tomando en cuenta que en Tablas de Contingencia los modelos están basados en independencia de probabilidades, de manera natural se consideran efectos multiplicativos y éstos son expresados por una liga logarítmica

$\eta = \log \mu$ (los efectos aditivos en η son

con inversa $\mu = e^{\eta}$ efectos multiplicativos en μ)

De esta manera, de acuerdo a los Modelos Lineales Generalizados se puede dar el siguiente ejemplo: Supóngase que las variables dependientes pueden ser consideradas como variables que tienen distribución Poisson con medias μ_i, esto es:

$$y_i = \mu_i + \varepsilon_i$$

sujetos a la condición que $\sum_{i=m}$ son fijos. Los componentes aleatorios ε_i representan las desviaciones aleatorias Poisson de la esperanza, mientras μ_i son los componentes sistemáticos.

Para una Tabla de Contingencia de 2 criterios, bajo la hipótesis de independencia entre renglones y columnas, donde y_i ocurre en los niveles k y l respectivamente a cada criterio de clasificación, se tiene:

$$\mu_i = m F_k F_l$$

donde Pk y Pl son las probabilidades de que la observación esté en la celda k,l. Sea

$$\eta_i = \log \mu_i \qquad \text{entonces}$$

$$\eta_i = \log m + \log P_k + \log P_l$$

$$= \sum_j x_{ij} b_j$$

para alguna b_j y con x_{ij} tomando valores 0 ó 1. Así

$$y_i = \exp \left(\sum_j x_{ij} b_j \right) + \varepsilon_i$$

se observa que y_i está expresada como una función (exponencial) de una suma lineal de componentes sistemáticos y un componente aleatorio.

3.2 Bondad de ajuste

La bondad de ajuste del modelo se determina con la devianza que es la medida de discrepancia establecido por el logaritmo de una razón de verosimilitudes.

Si tenemos N observaciones, el llamado modelo saturado tiene N parámetros linealmente independientes uno para cada observación además no tiene componente aleatorio (ya que toda la variación de las y's es explicada por el componente sistemático), reduciendo la variación residual igual a cero y un modelo de este tipo no es

el que se busca. El otro caso extremo es el del modelo nulo, el cual es equivalente a ajustar la media general de todas las observaciones y toda la variación en los datos es explicada por el componente aleatorio.

La devianza indica la discrepancia que existe cuando se ajustan modelos con un número menor de parámetros dado el modelo saturado. Para cada distribución tiene diferentes formas, para la distribución Poisson, que es de nuestro interés, se define como

$$\text{devianza} = 2[\sum y \ln(y/\hat{m})]$$

que precisamente es la estadística G^2 señalada en la sección anterior.

Los grados de libertad asociados a la devianza es la diferencia entre el No. de parámetros del modelo saturado y el No. de parámetros del modelo ajustado.

SEGUNDA PARTE

Una de las herramientas más útiles para el análisis de Tablas de Contingencia es el uso de paquetes estadísticos implementados en las computadoras.

Esta parte se divide en tres capítulos, correspondientes a los los paquetes estadísticos GLIM (Generalized Linear Models), SYSTAT (System for Statistics), BMDP (Biomedical Package). Estos permiten de una manera muy rápida en general, (dada la gran cantidad de operaciones) efectuar los cálculos para el ajuste de los modelos, aunque como se verá en la tabla final, depende del paquete estadístico que se utilice y el tipo de aplicación que el investigador le dé. En cada uno de éstos, se exhibe una breve descripción del paquete, cómo se accesa el sistema y la salida del mismo, la manera de definir y capturar los datos en el formato que se haya declarado para hacer la tabulación de los datos; y ya que se tiene un modelo, cómo declararlo en el paquete estadístico que se esté manejando y el ajuste del mismo. Asimismo se señala como obtener resultados del ajuste del modelo y las ventajas y desventajas que cada paquete ofrece.

El hecho de dar una estructura general de cada uno de los programas al final de cada manual, lo convierte en alguna fuente de referencia útil para dar una visión global de éstos.

Todas estas aplicaciones se muestran proporcionando una breve explicación de cada instrucción , la sintáxis y un ejemplo. Para hacer implícito el uso de las intrucciones, se incluirá un ejemplo que se marcará con 'ejm #', que tomará como base los

datos de la Tabla 2.3-1 y el modelo 2.3.19, expuestos en la I Parte de esta Tesis. Posteriormente, se integrarán las instrucciones para dar un ejemplo concreto del programa en cada paquete.

El estudio para los paquetes estadísticos GLIM y SYSTAT fueron efectuados en una microcomputadora BPM/51, aunque para la utilización de estos paquetes es suficiente contar con un microcomputadora de 256 Kb. Para el paquete BMDP se utilizó la computadora VAX.

1.GLIM

1.1 El paquete

GLIM es un programa en FORTRAN diseñado para realizar el ajuste de Modelos Lineales Generalizados. Se puede usar para modelos ANOVA, Tablas de Contingencia, Análisis Probit, etc.

Consiste en una sucesión de enunciados en la forma de directivos o comandos.

Un directivo consiste en un "nombre de directivo" (palabra reservada que se inicia con los símbolos $ ó \) seguido de un conjunto de campos y termina con los mismos símbolos $ ó \.

Los identificadores de más de cuatro caracteres se pueden usar, pero GLIM solamente toma en cuenta los 4 primeros caracteres (incluyendo $ ó \).

Ejm $DATA x,y$

1.2 Acceso al sistema

Antes de accesar el sistema debemos asegurarnos que en el diskette se cuenta con los programas : GLIMPROG.EXE, GLIMFLOP.BAT, GLIMWIC.BAT.

Posteriormente , para poder accesar el sistema GLIM, primero se deberá cargar el sistema MS-DOS, en seguida se insertará el sistema GLIM y aparecerá la señal:

A >

entonces teclear

A > GLIMPROG

el sistema responderá:

?GLIM ....

$UNITS

se teclearán el nombre del archivo de datos en el que se guardará el resultado del proceso que se realice.

ejm $UNITS PR1

Es conveniente guardar estos archivos en un segundo diskette, que se insertará en el drive B, se deberá teclear :

$UNITS b:<nombre del archivo>

En caso de que el nombre del archivo ya exista en el diskette, borrará el anterior y creará uno nuevo.

De la misma manera se procede para ?UNITS :

$UNITS b:<nombre del archivo>

1.3 Declaración y entrada de datos

Para definir el número (n) de celdas de la Tabla de Contingencia se emplea el comando UNITS.

Sintáxis: $UNITS n$

ejm * $UNITS 20$

$FACTOR

Se utiliza para declarar la dimensión de la tabla.

Sintáxis: $FACTOR <var1> #1 <var2> #2 ...$

donde : #n es un entero que indica el número de niveles de la variable.

ejm * $FACTOR PLANTEL 5 OCUP 2 INGRESO 2$

$CALC

Permite hacer cálculos en vectores. Es conveniente utilizarlo para declarar los índices de las celdas.

Sintáxis:

$CALC <var1> = %GL(Kn ,Nn) : <var2> = %GL(Kn ,Nn):...$

donde : Kn y Nn son enteros positivos y se genera un vector con números de 1 a Kn en bloques de Nn.

ejm *

$CALC PLANTEL = %GL(5,4): OCUP = %GL(2,1): INGRESO = %GL(2,2)$

genera los siguientes vectores (que indican las posiciones de las celdas).

plantel	ocup	ingreso
1	1	1
1	2	1
1	1	2
1	2	2
.	.	.
.	.	.
.	.	.
5	2	2

$DATA

Para indicar cuales son las variables que se van a utilizar (y que serán leídas con la instrucción $READ).

Sintáxis: $DATA <var>$

ejm1 * en caso de que anteriormente se haya incluído $CALC, para señalar los índices de las celdas, sólo quedaría leer las variables de las frecuencias.

$DATA FREC$

ejm2 * si no se utilizó $CALC anteriormente, se tienen que declarar las variables que indican los índices y las frecuencias de las celdas, con la desventaja de que es más tedioso el indicar los índices uno por uno.

$DATA PLANTEL OCUP INGRESO FREC$

La secuencia de la declaración de estas variables es de acuerdo al orden de los índices.

$READ

Los datos son leídos directamente de la terminal con este directivo.

Sintáxis: $READ [lista de datos] $

ejm * si se hizo la declaración como en ejm1 * de $DATA

$READ 18 820 37 1567 22 628 30 1190 21
855 38 1809 21 887 29 1412 33 901
60 23123$

(la secuencia de estos datos es de acuerdo a como se declaren los índices en el comando $CALC. En este caso tiene la secuencia del ejemplo que se dió en el comando $CALC anteriormente).

si se hizo la declaración como en ejm2 * de $DATA

$READ 1 1 1 18
1 2 1 820
.
.
.
5 2 2 23123$

Para esta instrucción es importante aclarar que el número mayor que admite el sistema es 32759.

$DINPUT

Con este comando, los datos pueden ser leídos directamente con un editor (p.e. Word Star), éste es muy útil ya que evita teclear los datos en cada ejecución del programa GLIM.

Sintáxis: $DINPUT <CANAL> <LONGITUD>$

donde : <canal> es un número entero entre 0 y 240, excepto los que se marquen en el directivo $ENVIRONMENT C que son los que ocupan los canales de entrada y salida propios del sistema. El número que se indique como canal es donde se va a direccionar el archivo que contiene los datos.

<longitud> es un número entero que determina la longitud del archivo de datos, este parámetro es opcional.

ejm al teclear

$DINPUT 9 30$

el sistema responderá:

File name ?

en este momento se deberá teclear el nombre físico del archivo de datos

ejm File name? b:datos.dat

1.4 Definición del modelo

Ya que se declaró la forma en que están organizados los datos, se procede a definir el modelo para posteriormente efectuar el ajuste deseado.

$YVAR

Para definir el modelo, es necesario declarar la variable

dependiente, para ésto se tecleará

Sintáxis: $YVAR <identificador>$
ejm * $YVAR FREC$

$ERR

La estructura del error, que en el caso de modelos loglineales es Poisson, se declara de la manera siguiente

Sintáxis: $ERR F$

donde F = POISSON

$LINK

Con este comando se declara la función liga que en el caso de modelos loglineales es la logaritmica (L)

Sintáxis: $LINK L$

1.5 Ajuste del modelo

$FIT

Después de declarar la variable dependiente, la distribución del error y la función liga, con el comando $FIT se procede al ajuste del modelo deseado.

Sintáxis: $FIT <modelo-fórmula>$

En el <modelo-fórmula> se declara el modelo que se requiera ajustar. Como resultado aparecerá en la pantalla el valor de la devianza y los grados de libertad para el modelo indicado.

ejm * para ajustar el modelo
PLANTEL+ OCUP+INGRESO+PLANTEL.INGRESO
la forma de declararlo es
$FIT %GM+PLANTEL+ OCUP+INGRESO+PLANTEL.INGRESO$

donde

%GM es la media general

PLANTEL, OCUP, INGRESO — representan desviaciones de la media general.

PLANTEL.INGRESO — es la asociación de 1' orden

Como resultado aparecerá

Scaled deviance = 8.6619 at cycle 3
d.f. = 9

Cuando se desea ajustar parámetro por parámetro se utiliza el símbolo ":"

ejm # $FIT %GM:+PLANTEL:+ OCUP:+INGRESO:+PLANTEL.INGRESO$

en este caso primero ajusta PLANTEL y calcula la devianza, después ajusta PLANTEL+OCUP y calcula la devianza y así sucesivamente hasta ajustar todo el modelo.

1.6 Desplegar y guardar resultados

$DISPLAY

Después del ajuste de datos, se requiere desplegar datos como los valores esperados estimados, los residuales,etc. estos son algunos resultados que se pueden obtener con este comando.

Sintáxis : $DISPLAY <letras>$

donde : <letras> puede ser un conjunto, en cualquier orden de las siguientes opciones:

E : Estimación de parámetros. Los estimadores de los parámetros son listados conjuntamente con sus errores estandar y los nombres de los parámetros .

R : Lista en paralelo los datos de los valores ajustados del modelo y los residuales estandarizados.

V : Lista las covarianzas de los parámetros estimados.

C : Lista las correlaciones de los parámetros estimados.

S : Lista los errores estandar de las diferencias de los parámetros estimados.

L : Lista el predictor lineal como una suma de términos.

M : Lista el modelo actual, incluyendo la variable dependiente, el error, la función liga.

A : Lista los parámetros estimados, los errores estándar y los nombres de parámetros en el mismo formato que la opción E, excepto que todos los parámetros son listados.

D : Lista la devianza y los grados de libertad del último modelo ajustado.

ejm * $DISPLAY R M1

Como resultado aparecerá

unit	observed	fitted	residual
1	18	21.36	-0.727
2	820	816.64	0.118
3	37	40.89	-0.608
4	1567	1563.11	0.098
5	22	16.57	1.354
6	628	633.43	-0.216
7	30	31.10	-0.197
8	1190	1188.90	0.032
9	21	22.33	-0.281
10	855	853.67	0.046
11	38	47.08	-1.324
12	1809	1799.92	0.214
13	21	17.44	0.854
14	663	666.56	-0.138
15	38	36.96	0.171
16	1412	1412.04	-0.078
17	33	23.81	1.804
18	901	910.19	-0.305
19	60	60.46	-0.060
20	2312	2311.54	0.010

```
current model:
number of units is 20
y-variate              FREC
weight                    *
offset                    *
probability distribution is POISSON
          link function is LOGARITHM
       scaled parameter is 1.000
terms = 1 + PLAN + OCUP + INGR + PLAN.INGR
```

$LOOK

Todos los ajustes proporcionan la siguiente información :

%DF - grados de libertad de la devianza

%DV - el valor de la devianza

%X2 - la estadística X^2 de Pearson

Para obtener cualquiera de éstas teclear:

Sintáxis: $LOOK <opción>$

ejm * $LOOK %DF$

como resultado aparecerá

9.000

Por último para terminar la sesión en un programa en GLIM, teclear:

$STOP

1.7 Facilidades del paquete GLIM

GLIM ofrece facilidades para su manejo, entre las más frecuentes se encuentran:

$ENVIRONMENT

Indica las características del paquete en la

microcomputadora como las que se señalan en las siguientes opciones:

C : Esta opción indica los números de los canales de entrada/salida

I : Proporciona detalles de la información, la representación de caracteres especiales y la longitud con que se pueden manejar los enteros.

D : Esta opción lista el espacio que internamente ocupan ciertos arreglos creados por la fórmula del modelo.

U : Señala las limitaciones en cuanto a espacio para los datos, el número de identificadores en el directorio, el número de vectores permitidos en la fórmula del modelo.

ejm $ENVIRONMENT C I4

$ACCURACY

Los números desplegados por el directivo $LOOK y $DISPLAY son dados por default con 4 dígitos significativos, que son redondeados, este número puede ser cambiado con el directivo $ACCURACY.

Sintáxis: $ACCURACY <no.entero>$

donde:<no.entero> son los dígitos significativos significativos. El valor máximo que permite es 9. El valor por default es 4.

$COMMENT

Con el este directivo se pueden escribir comentarios en el transcurso del proceso (estos no son ejecutables).

ejm $COMMENT ESTE ES UN COMENTARIO$

$REINPUT

En GLIM se puede trabajar en forma interactiva o en batch.

En forma interactiva es del tipo "pregunta-respuesta"; la forma batch consiste en elaborar un archivo con cualquier editor que contengan los directivos o instrucciones que se deseen trabajar y posteriormente enlazarla con el directivo $REINPUT.

Sintáxis: $REINPUT <integer1> <integer2>$

donde : <integer1> es el canal con el que se va a trabajar con el archivo externo. Es un número entero menor que 240 (excepto los que se indican con el comando $ENVIRONMENT C$).

<integer2> es el ancho de los renglones (es opcional).

ejm Supóngase que se tiene el archivo pr.dat que contiene los directivos requeridos, al teclear

$REINPUT 20 80$
aparecerá: File name?

en ese momento se tecleará el nombre del archivo pr.dat y se empezaran a ejecutar las instrucciones contenidas en este.

$ECHO

Cuando se emplea el comando $REINPUT, a veces es útil el saber que instrucciones está ejecutando, para esto se deberá teclear el comando $ECHO antes que $REINPUT, o al inicio del archivo externo. De esta forma aparecerá el nombre del del comando que se está ejecutando.

$LOOK

Este directivo permite ver los valores de la variable

especificada.

Sintáxis: $LOOK <identificadores>

ejm : $LOOK PLANTEL OCUP INGRESO FREC$

Como resultado aparecerá

```
 1   1   1    18
 1   2   1   820
     .
     .
     .
 5   2   2  2312
```

$CALC

GLIM permite hacer cálculos en vectores y escalares.

Sintáxis: $CALC <identificador-expresión>

ejm : $CALC X = 1$ (asigna 1a todos los elementos del vector X)

$CALC X=X + 1$ (a X le suma 1 modificando su valor inicial)

$DUMP y $RESTORE

Cuando se está trabajando en forma interactiva y se desea suspender la ejecución para después continuarlo almacenando los resultados ya obtenidos se emplea $DUMP. Esta instrucción genera un archivo en código el cual respalda todas las instrucciones y resultados ya generados.

Sintáxis: $DUMP <integer>$

donde
<integer> es el canal en el que va a ser guardado el archivo.

ejm :

```
$DUMP 77$
File name : b:resp
```

y estará respaldado cuando el sistema contesta

```
--program dump completed
```

Para restablecer el programa, en seguida que se accesa GLIM la primera instrucción será con $RESTORE

Sintáxis: $RESTORE <integer>$

donde <integer> es el canal indicado en $DUMP

ejm $RESTORE 77$

File name : b:resp

1.8 ESTRUCTURA GENERAL DE UN PROGRAMA EN GLIM

```
+ *     $ECHO
+ *     $COMMENT
+ *     $ENVIRONMENT
  *     $REINPUT
        $UNITS
        $FACT
        $CALC
+ *     $ACCURACY
        $DATA
+ *     $LOOK
        $YVAR
        $READ o $DINPUT
        $ERROR
        $LINK
+ #     $FIT
+ #     $DISPLAY
+ #     $LOOK   (%DF, %PV, %X2)
        $STOP
```

Notas

* Opcional

\+ Se puede utilizar a partir del lugar indicado, en cualquier parte del programa, sin modificar la estructura de la declaración de las variables.

\# Se puede utilizar repetidas veces tomando en cuenta los comandos declarados anteriormente.

EJEMPLO DE UN PROGRAMA EN GLIM:

```
GLIM 3.77 update 0 (copyright)1985 Royal Statistical Society, Lo

? $UNITS 20$

? $FACTOR PLANTEL 5 OCUP 2 INGRESO 2$

? $CALC PLANTEL=%GL(5,4): OCUP=%GL(2,1): INGRESO=%GL(2,2)$

? $DATA FREC$

? $READ 18 820 37 1567 22 628 30 1190 21 855 38 1809 21
$REA? 663 38 1412 33 901 60 2312$

? $LOOK PLANTEL OCUP INGRESO FREC$
      PLAN     OCUP     INGR      FREC
   1  1.000    1.000    1.000     18.00
   2  1.000    2.000    1.000    820.00
   3  1.000    1.000    2.000     37.00
   4  1.000    2.000    2.000   1567.00
   5  2.000    1.000    1.000     22.00
   6  2.000    2.000    1.000    628.00
   7  2.000    1.000    2.000     30.00
   8  2.000    2.000    2.000   1190.00
   9  3.000    1.000    1.000     21.00
  10  3.000    2.000    1.000    855.00
  11  3.000    1.000    2.000     38.00
  12  3.000    2.000    2.000   1809.00
  13  4.000    1.000    1.000     21.00
  14  4.000    2.000    1.000    663.00
  15  4.000    1.000    2.000     33.00
  16  4.000    2.000    2.000   1412.00
  17  5.000    1.000    1.000     33.00
  18  5.000    2.000    1.000    901.00
  19  5.000    1.000    2.000     60.00
  20  5.000    2.000    2.000   2312.00

? $YVAR FREC$

? $ERR P$

? $LINK L$

? $FIT %GM+PLANTEL+OCUP+INGRESO+PLANTEL.INGRESO$
scaled deviance = 8.6619 at cycle  3
            d.f. = 9

? $DISPLAY R M$
  unit  observed     fitted   residual
     1        18      21.36     -0.727
     2       820     816.64      0.110
     3        37      40.89     -0.608
     4      1567    1563.11      0.098
     5        22      16.57      1.334
     6       628     628.45     -0.216
     7        30      31.10     -0.197
```

```
      8       1190     1188.90        0.032
      9         21       22.33       -0.291
     10        855      853.67        0.046
     11         38       47.08       -1.324
     12       1809     1799.92        0.214
     13         21       17.44        0.854
     14        667      666.56       -0.138
     15         38       36.96        0.171
     16       1412     1413.04       -0.028
     17         33       23.81        1.884
     18        901      910.19       -0.305
     19         60       60.46       -0.060
     20       2312     2311.54        0.010

Current model:

  number of units is  20

  y-variate  FREC
  weight     *
  offset     *

  probability distribution is POISSON
            link function is LOGARITHM
         scale parameter is 1.000

  terms =  1 + PLAN + OCUP + INGR + PLAN.INGR

? $LOOK %DF %X2$
   9.000   9.089

? $STOP$
```

2. SYSTAT

2.1 El paquete

SYSTAT es un paquete estadístico que contiene varios módulos para el ajuste y el análisis de datos como son STATS (Estadística Univariada), TABLES (Tablas de Contingencia y Modelos Loglineales), CORR (Coeficientes de Correlación), MGLH (Hipótesis Lineal General Multivariado), FACTOR (Análisis de Componentes Principales), MDS (Escalamiento Multidimensional), CLUSTER (Análisis Cluster), NPAR (Estadística no Paramétrica), SERIES (Análisis de Series de Tiempo), así como un módulo de graficación y otro para la captura de datos.

El módulo de interés es TABLES. Este módulo consiste en un conjunto de comandos que se pueden dar en forma interactiva o en forma batch.

Los dos principales comandos en el módulo: TABULATE y MODEL pueden ser usados repetidamente con los mismos datos para las tabulaciones y la declaración de los modelos.

2.2 Acceso al sistema

El proceso para el ajuste de datos en SYSTAT consta de dos

partes: una para el registro y captura de datos, que es con el módulo DATA, y la otra es para el ajuste del los datos usando el módulo TABLES.

Antes de accesar el sistema se debe asegurar que en los diskettes se encuentren los siguientes programas:

TABLES	{	TABLES.EXE TABLES.DEF
DATA	{	DATA.EXE DATA.DEF

Para accesar cualquiera de los dos módulos primero se deberá cargar el sistema operativo MS DOS y aparecerá la siguiente señal:

```
A>
```

En seguida se insertará el segundo diskette que contenga el módulo que se desea trabajar (ya sea DATA o TABLES) y teclear:

```
A> DATA (TABLES)
```

el sistema responderá:

```
---------
---------------
SYSTAT
--------------
         ---------

VERSION 2.1
COPYRIGHT, 1985
SYSTAT, INC
SERIAL NUMBER IS:
YOU ARE IN DATA (TABLES) MODULE
WORKSPACE CLEAR FOR CREATING NEW DATA SET
```

En este momento ya está lista para empezar a trabajar.

Para abandonar el sistema teclear:

>QUIT

2.3 Declaración y entrada de datos

Para la captura y registro de datos se usará el módulo DATA.

En SYSTAT se pueden registrar los datos directamente tecleándolos con el editor que contiene SYSTAT que es parecida a una hoja de cálculo, con la desventaja que no se pueden hacer cálculos y por otra parte resulta muy lento el proceso de captura de datos, o mediante un archivo previamente creado por paquetes como DBASE, WORDSTAR, LOTUS1-2-3 y posteriormente transformarlos a un archivo de SYSTAT y después accesarlos con el módulo TABLES para efectuar el ajuste.

Es importante señalar que los números deberán ser de 9 dígitos máximo.

A continuación se muestran 3 formas de registrar los datos:

1° Usando el editor de SYSTAT

1° EDIT

Para registrar los datos directamente con el editor de SYSTAT, teclear lo siguiente:

A>DATA
>EDIT

En este momento se podrá empezar a teclear los datos en las columnas correspondientes, para iniciar se deben teclear los nombres de las variables en el primer

renglón, el nombre debe empezar con un apóstrofe. En los siguientes renglones se capturan los datos, moviendo el cursor a columnas y renglones con las flechas que contiene el teclado.

Las opciones que tiene el editor son:

(para utilizarlas primero dar <escape> o O)

FIND <expresión>: mueve el cursor a partir de donde se encuentra al caso o variable seleccionada.
ejm >FIND CASE=5

FORMAT <#>: cambia el formato de los números.
ejm >FORMAT =0 (los números los despliega sin decimales).

SAVE <nomarch>: salva el nombre del archivo.

IF <expresión> THEN LET <oración> :
si se cumple <expresión> entonces se ejecuta la <oración>.

HELP informa de los comandos del editor y del movimiento del cursor.

END termina el modo de edición.

Ejm $

```
>EDIT
 (aparece la ventana de edición)
 SYSTAT Editor
 ------------------------------------------------------
 case¦
   1 ¦
   2 ¦
   3 ¦
   . ¦
   . ¦
   . ¦
```

En este momento se pueden empezar a capturar las variables de la siguiente forma:

'PLANTEL	'OCUP	'INGRESO	'FREC
1	1	1	18
1	2	1	820
	.		
	.		
	.		
5	2	2	2312

teclear <escape> ó 0

>SAVE PRUEBA (al salvar el archivo, tendrá la terminación .SYS)

>END

termina el modo de edición con el siguiente mensaje:

>WORKSPACE CLEAR FOR CREATING NEW DATA SET

En caso de querer accesar un archivo ya creado, teclear :

Sintáxis >EDIT <nom arch>

ejm >EDIT PRUEBA

(edita el archivo PRUEBA.SYS)

2' Directo en forma interactiva

Otra forma de capturar los datos en el módulo DATA, es interactivamente con el siguiente procedimiento:

1' SAVE

Este comando salva los datos en un archivo SYSTAT. Para crear un archivo se deberá usar este comando antes del comando RUN.

Sintáxis: >SAVE <nom arch>

2' INPUT

Con este comando se especifica el nombre de las variables el cual deberán empezar con una letra y ser de 8 caracteres máximo.

Sintáxis : >INPUT VAR1 VAR2 ... <formato>

donde <formato> deberá empezar con el signo $ si la variable es alfabética.

3' RUN

Este comando 'recoge' los datos de la consola y efectúa el proceso.

Sintáxis: >RUN

(teclear los datos)

4' NEW

Limpia el espacio de trabajo como si se empezara a trabajar.

Sintáxis: >NEW

Ejem *

Supóngase que se registrarán los datos en un archivo llamado PRUEBA:

```
>SAVE PRUEBA
>INPUT PLANTEL  OCUP INGRESO FREC
>RUN
INPUT DATA ONE CASE AT TIME AFTER PROMPT ARROW
 1 1 1     18
 1 2 1    820
 1 1 2     37
 1 2 2   1567              ( A )
 2 1 1     22
 2 2 1    628
 2 1 2     30
 2 2 2   1190
 3 1 1     21
 3 2 1    855
 3 1 2     38
 3 2 2   1809
 4 1 1     21
 4 2 1    663
 4 1 2     30
 4 2 2   1412
 5 1 1     33
 5 2 1    901
 5 1 2     40
 5 2 2   2312
>NEW
      20 CASES AND 4 VARIABLES PROCESED
      SYSTAT FILE CREATED
      WORKSPACE FOR CRATING NEW DATA SET
```

El problema de este procedimiento es que el archivo que crea es en el código del sistema y en caso de querer corregir algún dato, no se puede efectuar directamente, se tienen que usar las expresiones IF y LET:

Sintáxis: >IF <expresión> THEN <oración>

En el ejemplo anterior, si se quiere cambiar de valor a 100 la celda 1,1,1 se hará de la siguiente forma:

```
>USE PRUEBA
>SAVE PRUEBA1
>IF CASE=1 THEN LET FREC=100
 ó >IF PLANTEL=1 AND OCUP=1 AND INGRESO=1 THEN LET FREC=100
>RUN
```

En este ejemplo se emplea el comando USE que se utiliza para leer los datos en un archivo SYSTAT.

Sintáxis:

>USE <arch1> [(<var1>,...)] <arch2> [(<var1>,...)] ...

3° A través de un archivo en forma batch

La forma más práctica de registrar los datos es usando un archivo externo, que se puede crear con paquetes como: LOTUS 1-2-3, DBASE, WORDSTAR que generan archivos en código ASCII .El nombre de este archivo deberá tener la terminación .DAT. Para poder trabajar con éste, se tiene primero que transformar en un archivo SYSTAT que se obtiene con el siguiente procedimiento:

```
>GET <nom arch>              (sin la terminación .DAT)
>INPUT <nom arch>
>SAVE <nom arch>
>RUN
Y se obtendrá el siguiente mensaje:
X CASES AND Y VARIABLES PROCESSED SYSTAT CREATED
WORKSPACE CLEAR FOR CREATING NEW DATA SET
>
```

Este mensaje indica que el archivo SYSTAT fué creado, y muestra el no. de casos (X) y no. de variables(Y) que se procesaron.

En este ejemplo se utilizó el comando GET que sirve para leer los datos en el archivo <nom arch> que en el directorio deberá aparecer con la extensión .DAT, es decir <nom arch.DAT>. los restantes comandos funcionan como se explicó anteriormente.

Ejemplo: Para registrar los datos de la tabla 2.3-1, señalada en la I Parte, en WORDSTAR se puede hacer de la siguiente manera:

1' A>WS
2' Teclear la opción N para abrir el archivo.
3' Teclear el nombre del archivo con extensión .DAT
ejm PRUEBA.DAT
4' Teclear los datos
(de la misma manera que en A).
5' Salvar el archivo y abandonar el sistema con ^KX
6' Cargar el módulo DATA:
A>DATA
7' Ya que se accesó este módulo, teclear:
>GET PRUEBA
>INPUT PLANTEL OCUP INGRESO FREC
>SAVE PRUEBA
>RUN

El sistema contestará:
20 CASES AND 4 VARIABLES PROCESSED
SYSTAT FILE CREATED
WORKSPACE FOR CREATING NEW DATA SET

En este momento ya se tiene el archivo PRUEBA ya transformado a SYSTAT.

LIST

Para verificar el registro de los datos, se pueden listar. Cualquier archivo SYSTAT que se haya generado con algún procedimiento señalado anteriormente se lista con este comando.

Sintáxis: >LIST*

ejm :para listar el archivo de SYSTAT llamado PRUEBA, teclear:

```
>USE PRUEBA
 (El sistema responderá)
  SYSTAT FILE VARIABLE AVAIBLE TO YOU ARE
                PLANTEL  OCUP INGRESO FREC
>LIST
>RUN
 (aparecerán los datos listados como en (A))
```

2.4 Tabulación de datos

Ya que se tiene listo el archivo de datos en SYSTAT, para efectuar el ajuste del modelo se requiere primero 'cargar' el archivo de datos (que tiene la terminación .SYS) con el comando USE mencionado anteriormente y de esta manera se podrán tabular los datos y especificar la Tabla de Contingencia con el siguiente comando:

TABULATE

Sintáxis:

```
>TABULATE <VAR1>*<VAR2>,... /WEIGHT=VAR3, FREQUENCY,
                              PERCENT,ROWPCT,MISS,LIST
```

donde:

WEIGHT: Tabula el valor de la variable especificada con la frecuencia correspondiente a la celda. Esta opción siempre es necesaria especificarla en el comando TABULATE.

FREQUENCY: Tabula las frecuencias de los valores observados de la tabla.

PERCENT: Tabula los porcentajes de cada subtabla.

ROWPTC: Tabula los porcentajes por renglón de cada subtabla.

COLPCT: Tabula los porcentajes por columna de cada subtabla.

MISS : Omite las entradas de los valores de la celda iguales a cero.

LIST : Lista en una tabla los valores acumulados de la variable señalada con la opción WEIGHT.

ejm *

```
>USE PRUEBA
>TABULATE PLANTEL* OCUP*INGRESÓ/WEIGHT=FREC
```

Como resultado aparecerán 5 tablas con las tabulaciones para cada plantel, similares a la que a continuación se muestra:

```
TABLE OF OCUP (ROWS) BY INGRESO (COLUMNS)
FOR THE FOLLOWING VALUES:

        PLANTEL = 1

FRECUENCIES

        1       2      TOTAL
     -------------
1 :    18      37 :      55
2 :   820    1567 :    2387
     -------------
```

PRINT

Si se desea obtener algún coeficiente de asociación como la X^2 de Yates, coeficiente de contingencia, Q de Yules y la Y de Yules, que se calculan sólo para tablas de 2 criterios, se tendrá que especificar la opción:

```
>PRINT LONG
```

que deberá declararse antes de >TABULATE.

La opción por default es:

```
>PRINT SHORT
```

de esta manera no se calculan los coeficientes de asociación.

2.5 Definición, ajuste del modelo y obtención de resultados

Después que se hizo la tabulación se puede efectuar el ajuste con el siguiente comando:

MODEL

Sintáxis:

```
>MODEL <VAR1> + <VAR2> +....+<VAR1>*<VAR2>+.....+
       <VAR1>*<VAR2>*<VAR3>+..../ITERATION<#>,DELTA=<#>,
                          FITTED,DIFFERENCES,RESIDUALS,ZERO
```

donde:

DELTA=<#> : con esta opción se suma la cantidad indicada con <#> a cada celda antes de hacer el ajuste.

FITTED : despliega los valores ajustados.

DIFFERENCES: calcula las diferencias de valores esperados y ajustados.

RESIDUALS : despliega los residuales.

ZERO : los valores de las celdas iguales a cero no son ajustadas. Los toma como cero estructural, i.e. que probablemente en esa celda el valor observado sea 0. Los valores observados iguales a cero tendrán valores esperados iguales a cero.

ITERATION<#>: limita el número de iteraciones al especificado en #.

CRITERION : modifica el criterio para terminar las iteraciones.

Como resultado del comando MODEL, se obtiene el número de interacciones, el cálculo de la X^2, los grados de de libertad, la probabilidad y la G^2.

ejm *

Para ajustar el modelo donde hay asociación entre

PLANTEL e INGRESO

>MODEL PLANTEL+ OCUP+INGRESO+PLANTEL*INGRESO

Como resultado aparecerá

```
MODEL WAS FIT AFTER 2 ITERATIONS.
TEST OF FIT MODEL
DEGREES OF FREEDOM = 9
PEARSON CHI-SQUARE = 9.09   PROBABILITY = .429
LIKELIHOOD RATIO CHI-SQUARE = 8.66   PROBABILITY = .469
```

Salida del sistema

Para abandonar el sistema, bastará con teclear

>QUIT

(en este momento aparecerá una lista con todos los comandos tecleados y el prompt de MS DOS)

2.6 Facilidades del paquete SYSTAT

SYSTAT ofrece facilidades para su manejo como las que se señalan a continuación:

HELP

Informa de un comando en particular o mas comandos que existen en el módulo, indicando la sintáxis de cada uno.

Sintáxis: >HELP
>HELP <comando>

SUBMIT

En SYSTAT se puede trabajar en forma interactiva o batch. La forma interactiva es el tipo "pregunta-

respuesta" la forma batch consiste en elaborar un archivo que contenga los comandos que se desean emplear y posteriormente accesarlo. Este archivo deberá estar en el directorio de archivos y tener la terminación .CMD.

Sintáxis: >SUBMIT <arch>

Ejm >SUBMIT PRUEBA (lee y ejecuta los comandos tecleados en PRUEBA.CMD)

OUTPUT

Este comando direcciona la salida de información a la consola, a un archivo en disco o a la impresora. Se puede usar en todo el proceso o en parte de él.

Sintáxis:

```
>OUTPUT *       (envía los resultados a la pantalla)
>OUTPUT @       (envía los resultados a la impresora)
>OUTPUT <arch>  (envía los resultados a un archivo
                 ejm  >OUTPUT SALIDA
                      los envía al archivo  SALIDA.DAT).
```

NOTE

Con este comando se pueden escribir comentarios.

Sintáxis: >NOTE 'comentario'

Ejm >NOTE 'este es un comentario'

PAGE

Este comando limita el formato de la salida a 80 ó 132 columnas.

```
Sintáxis:   >PAGE NARROW   (80 columnas)
            >PAGE WIDE     (132 columnas)
```

SELECT (sólo para el módulo TABLES)

Selecciona casos de un archivo para el análisis que se

vaya a realizar.

Sintáxis: >SELECT <oración>
Ejm >SELECT PLANTEL=1 OCUP=1 INGRESO=1

sólo tomará en cuenta para el análisis la celda 1,1,1.

FORMAT

El formato de los números que aparecen en los marginales de las tablas, puede ser cambiado con este comando. Determina el número de dígitos (que no debe ser mayor que 9) a la derecha del punto decimal. El valor que toma por default es 3.

Sintáxis: >FORMAT <no. dígitos>

Ejm >FORMAT 5 (toma número de la forma 6.00000)

DOS

Con este comando se dá acceso a DOS y ejecuta cualquier comando de DOS.

Sintáxis: >DOS <comando>

Ejm >DOS DIR
>DOS 'DEL PRUEBA.DAT'

2.7 ESTRUCTURA GENERAL DE UN PROGRAMA EN SYSTAT

Módulo DATA

For editor	Forma interactiva	Forma batch
* HELP	* HELP	* HELP
* OUTPUT	* OUTPUT	* SUBMIT
* NOTE	* NOTE	* OUTPUT
* PAGE	* PAGE	* NOTE
* DOS	* DOS	* PAGE
EDIT	SAVE	GET
SAVE	INPUT	INPUT
QUIT	RUN	SAVE
	NEW	* LIST
	QUIT	RUN
		QUIT

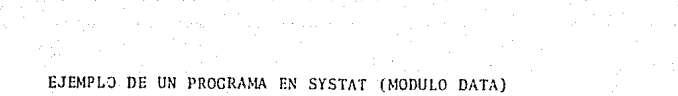
Módulo TABLES

* SUBMIT
* OUTPUT
* NOTE
* PAGE
* FORMAT
* PRINT
  USE
* SELECT
  TABULATE
  MODEL
* NEW
  QUIT

Notas:

* Opcional

EJEMPLO DE UN PROGRAMA EN SYSTAT (MODULO DATA)



VERSION 2.1
COPYRIGHT, 1985
SYSTAT, INC.
SERIAL NUMBER IS: 4165

THIS PROGRAM BELONGS TO BUSINESS MICROCOMPUTER SOLUTIONS, AUSTIN, TEXAS
PLEASE CALL (512)-864-5670 IF YOU FIND AN ILLEGAL COPY

YOU ARE IN DATA MODULE
WORKSPACE CLEAR FOR CREATING NEW DATASET
>

>
>SAVE A:PRUEBA
>INPUT PLANTEL CCUF INGRESO FREC
>RUN
INPUT DATA ONE CASE AT A TIME AFTER PROMPT ARROW
>1 1 1 18
>1 2 1 820
>1 1 2 37
>1 2 2 1567
>2 1 1 22
>2 2 1 628
>2 1 2 30
>2 2 2 1190
>3 1 1 21
>3 2 1 855
>3 1 2 38
>3 2 2 1809
>4 1 1 21
>4 2 1 665
>4 1 2 38
>4 2 2 1412
>5 1 1 32
>5 2 1 901
>5 1 2 60
>5 2 2 2312

>NEW

20 CASES AND 4 VARIABLES PROCESSED.
SYSTAT FILE CREATED.
WORKSPACE CLEAR FOR CREATING NEW DATASET
WORKSPACE CLEAR FOR CREATING NEW DATASET
>

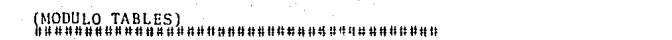
>
>QUIT

```
SYSTAT PROCESSING FINISHED

INPUT STATEMENTS FOR THIS JOB:


SAVE A:PRUEBA
INPUT PLANTEL OCUP INGRESO PREC
NEW
Stop - Program terminated.

C>
```




```
VERSION 2.1
COPYRIGHT. 1985
SYSTAT, INC.
SERIAL NUMBER IS: 4165

THIS PROGRAM BELONGS TO BUSINESS MICROCOMPUTER SOLUTIONS, AUSTIN, TEXAS
PLEASE CALL (512)-264-5670 IF YOU FIND AN ILLEGAL COPY

YOU ARE IN TABLES MODULE



>USE A:PRUEBA

VARIABLES IN SYSTAT FILE ARE:
   PLANTEL          OCUP          INGRESO          FREC

>TABULATE PLANTEL*OCUP*INGRESO / WEIGHT=FREC
 TABLE OF    OCUP     (ROWS) BY  INGRESO     (COLUMNS)
    FOR THE FOLLOWING VALUES:
         PLANTEL     =          1

 FREQUENCIES

                 1          2      TOTAL
          ----------------------
        1 |       18         37 |      55
          |                     |
        2 |      820       1567 |    2387
          ----------------------
 TOTAL            838       1604      2442


 TABLE OF    OCUP     (ROWS) BY  INGRESO     (COLUMNS)
    FOR THE FOLLOWING VALUES:
         PLANTEL     =          2

 FREQUENCIES

                 1          2      TOTAL
          ----------------------
        1 |       22         30 |      52
          |                     |
        2 |      628       1190 |    1818
          ----------------------
 TOTAL            650       1220      1870
```

```
TABLE OF      OCUP       (ROWS) BY  INGRESO        (COLUMNS)
   FOR THE FOLLOWING VALUES:
          PLANTEL       =           3

FREQUENCIES

                   1             2      TOTAL
          ------------------------------
       1 |        21            38 |      59
         |                         |
       2 |       855          1809 |    2664
          ------------------------------
TOTAL            876          1847      2723

TABLE OF      OCUP       (ROWS) BY  INGRESO        (COLUMNS)
   FOR THE FOLLOWING VALUES:
          PLANTEL       =           4

FREQUENCIES

                   1             2      TOTAL
          ------------------------------
       1 |        21            38 |      59
         |                         |
       2 |       663          1412 |    2075
          ------------------------------
TOTAL            684          1450      2134

TABLE OF      OCUP       (ROWS) BY  INGRESO        (COLUMNS)
   FOR THE FOLLOWING VALUES:
          PLANTEL       =           5

FREQUENCIES

                   1             2      TOTAL
          ------------------------------
       1 |        33            60 |      93
         |                         |
       2 |       901          2312 |    3213
          ------------------------------
TOTAL            934          2372      3306


>MODEL PLANTEL + OCUP + INGRESO + PLANTEL*INGRESO / FITTED

MODEL WAS FIT AFTER  2 ITERATIONS.
TABLE OF      OCUP       (ROWS) BY  INGRESO        (COLUMNS)
   FOR THE FOLLOWING VALUES:
          PLANTEL       =           1
```

FITTED VALUES

	1	2
1	21.36	40.89
2	816.64	1563.11

TABLE OF OCUP (ROWS) BY INGRESO (COLUMNS)
FOR THE FOLLOWING VALUES:
PLANTEL = 2

FITTED VALUES

	1	2
1	16.57	31.10
2	633.43	1188.90

TABLE OF OCUP (ROWS) BY INGRESO (COLUMNS)
FOR THE FOLLOWING VALUES:
PLANTEL = 3

FITTED VALUES

	1	2
1	22.33	47.08
2	853.67	1799.92

TABLE OF OCUP (ROWS) BY INGRESO (COLUMNS)
FOR THE FOLLOWING VALUES:
PLANTEL = 4

FITTED VALUES

	1	2
1	17.44	36.96
2	666.56	1413.04

TABLE OF OCUP (ROWS) BY INGRESO (COLUMNS)
FOR THE FOLLOWING VALUES:
PLANTEL = 5

```
FITTED VALUES

                1          2
        ---------------------
      1 |   23.81      60.46  |
        |                     |
      2 |  910.19    2311.54  |
        ---------------------


 TEST OF FIT OF MODEL

    DEGREES OF FREEDOM =    9
    PEARSON CHI-SQUARE =          9.09    PROBABILITY =   .429
    LIKELIHOOD RATIO CHI-SQUARE =         8.66    PROBABILITY =   .469

>QUIT
 SYSTAT PROCESSING FINISHED

 INPUT STATEMENTS FOR THIS JOB:


 USE A:PRUEBA
 TABULATE PLANTEL*OCUP*INGRESO / WEIGHT=FREC
 MODEL PLANTEL + OCUP + INGRESO + PLANTEL*INGRESO / FITTED
Stop - Program terminated.
```

BMDP

3.1 El paquete

BMDP es un conjunto de programas diseñado para el análisis de datos, usando desde métodos sencillos hasta técnicas estadísticas avanzadas. Los datos generalmente son analizados, al igual que otros paquetes, por un iterativo "examinar y modificar", esto es que primeramente se analizan los datos y después del análisis, se modifica el modelo y se vuelve a examinar.

Los programas que contiene el paquete estadístico BMDP son: Descripción de Datos, Tablas de Contingencia, Análisis de Regresión, Análisis de Varianza, Análisis Multivariado, Series de Tiempo. Cada programa se identifica por un código. El código correspondiente Tablas de Contingencia es P4F.

Un programa en BMDP consiste en un conjunto de enunciados (llamados Lenguaje de Control) que se inician con un slash, seguido de una o varias palabras reservadas y un conjunto de campos, y terminan con un punto. Ejm

/VARIABLE NAMES ARE INGRESO, OCUP.

Opcionalmente en el programa pueden ir incluídos los datos o pueden ser leídos de otra fuente .

Los programas en BMDP pueden ser ejecutados interactivamente o

en forma batch.

Para terminar un programa en BMDP, teclear: /END si fué ejecutado en forma batch o /FINISH si fué ejecutado en forma interactiva.

Es importante señalar que el programa P4F puede analizar tablas de dos criterios con un máximo de 5000 celdas y de n-criterios con un máximo de 3000 celdas.

3.2 Acceso y salida al y del sistema

Para hacer el análisis de datos en forma batch es necesario primeramente teclear los comandos requeridos con el editor de la computadora VAX y posteriormente ejecutar el programa. En la forma interactiva, se teclearán los comandos cada vez que el programa los requiera. En este trabajo solo se utilizará la forma batch, ya que tiene la ventaja de ahorrar tiempo.

Para accesar el editor de la VAX, primero se deberá teclear la clave del usuario y su password, posteriormente aparecerá la siguiente señal con la fecha del día que se accesa.

```
Last interactive login on Thursday, 27 aug 1987
```

Para crear un programa llamado por ejemplo PRUEBA.DAT, teclear:

```
$EDIT PRUEBA.DAT
```

en este momento se pueden empezar a teclear las instrucciones requeridas. Al terminar, se teclearé:

```
*EXIT
```

en este momento el programa se salva y ya puede ser ejecutado. Para ejecutar el programa correspondiente, teclear:

```
$BMDF P4F
```

y el sistema responderá:

```
Name of BMDP Instruction Languaje File:
```

teclear el nombre del programa. Ejm:

```
Name of BMDP Instruction Languaje File: PRUEBA.DAT
```

posteriormente teclear el nombre de salida:

```
Name of file to write output to: PRUEBA.SAL
```

en este momento el programa se empieza a ejecutar, apareciendo la señal:

```
Now running  4F ....
```

en caso de que no hubiera errores en el programa, el sistema finalmente preguntará:

```
Name of BMDP program to run : <dar return>
All done
$
```

Con estos mensajes termina la ejecución del programa y los resultados obtenidos se pueden ver en el archivo de salida (en este ejemplo, PRUEBA.SAL).

De una manera más fácil se puede ejecutar de la siguiente manera, que es equivalente a la anterior.

```
$BMDP P4F PRUEBA.DAT PRUEBA.SAL
```

3.3 Declaración de datos

La intrucciones usadas para describir los datos y las variables,

son:

/PROBLEM

Se utiliza para dar un título al análisis a ser efectuado, éste no debe exceder de 160 caracteres. Si se omite, el título de la salida quedará en blanco.

Sintáxis: /PROBLEM TITLE = 'enunciado'.

Ejm /PROBLEM TITLE = 'ESTE ES UN EJEMPLO'.

/INPUT

Describe la entrada de los datos (número de casos, número de variables y el formato). Para esto es necesario especificar cada caso.

VARIABLES

Indica el número de variables (#).

Sintáxis: VARIABLES = #.

FORMAT

Describe el formato en el que van a ser leídos los datos (se usa el mismo que utiliza el lenguaje FORTRAN), se puede usar también formato libre (FREE).

Sintáxis: FORMAT = '<FORMATO>'.

CASES

Indica el número de casos. Este no es necesario especificarlo a menos que sólo una parte se lean o cuando en una ejecución se hagan repetidos análisis.

Sintáxis: /CASE = <#>.

donde <#> es el número de casos.

ejm # del comando /INPUT:

```
/INPUT    VARIABLES = 3.
          CASES = 20.
          FORMAT = FREE.
```

/VARIABLE

Sirve para declarar las variables.

NAME

Declara las variables, el nombre de éstas debe ser menor de 9 caracteres.

Sintáxis: NAME <lista de variables>.

Si el nombre de la variable no comienza con alguna letra, éste deberá estar entre apóstrofes.

Ejm * :

/VARIABLE NAME = PLANTEL, OCUP,INGRESO.

/CATEGORY (o GRUP)

Es usado para clasificar los casos en grupos o categorías.

CODE y NAME

Son utilizados para identificar las categorías de las variables.

Sintáxis: CODE (#) = #lista.

NAME (#) = nombre lista. El nombre debe ser menor de 9 caracteres y en caso de que no comience con letra, deberá estar entre apóstrofes.

CUTPOINTS

Se usa en lugar de CODE para separar una variable(continua) en intervalos.

Sintáxis: CUTPOINTS (#) = #lista.

Ejm * considerando una tabla marginal:

	INGRESO	
PLANTEL	primer ingreso	reingreso
azc	838	1604
nau	650	1220
ori	876	1847
sur	684	1450
val	934	2372

la declaración será:

```
/VARIABLE NAMES = PLANTEL,INGRESO.
/CATEGORY CODES (1) = 1,2,3,4,5.
          NAMES (1) = azc,nau,ori,sur,val.
          CODES (2) = 1,2.
          NAMES (2) = 'pri ing','reing'.
```

En este sentido, CODES (1) = 1,2,3,4,5. Significa que los códigos de la variable PLANTEL van a ser 1,2,3,4,5 y los va a llamar azc,nau,ori,sur,val respectivamente. De esta manera, un individuo que tenga los códigos 1,1,1 quiere decir que es del plantel azcapotzalco, de primer ingreso y es patrón.

Es importante señalar que CATEGORY CODES o CUTPOINTS son requeridos si hay más de 10 valores para una variable categórica o si se tiene una tabla con n-criterios; de otra manera 10 niveles son ocupados por cada índice y en este caso se puede exceder el espacio disponible en la memoria de la computadora. En algunos casos donde ya existen métodos para tabulación de datos esta instrucción es omitida.

/RESET

Cuando esta instrucción es usada, todas las asignaciones efectuadas son canceladas.

3.4 Tabulación de datos

Para hacer la tabulación de datos en BMDP existen varias formas, algunas para tablas de 2 criterios y otras para tablas de 3 ó más criterios, todas usando la instrucción TABLE (que se hace referencia mas adelante), además se tienen las opciones PRINT, STATISTICS, que son opcionales de acuerdo al análisis requerido.

3.4.1 Tablas de dos criterios

/TABLE

Con este comando podemos tabular los datos de una o varias tablas de 2 criterios. Para esto se tiene que indicar que categorías forman las columnas y cuales los renglones.

Los comandos opcionales son PRINT, STATISTICS, DELTA.

Sintáxis: /TABLE COL = <lista var>. (define las categorías de las columnas)
ROW = <lista var>. (define las categorías de los renglones)

PAIR o CROSS. (las tablas son formadas de todas las posibles combinaciones de COL y ROW)

Ejemplo para tabular los datos:

```
1 1 1 1
1 1 1 1
1 2 1 2
1 2 2 2
2 1 2 1
2 2 2 2
2 1 2 1
2 1 1 2
2 2 2 2
2 1 2 1
1 2 1 2
1 1 2 2
```

con la instrucción

/TABLE COL = EDAD,OCUP,EDAD.
ROW = PLANTEL,PESO,OCUP.

Se obtendrá como resultado las siguientes tablas y resultados:

	edad				ocup				edad			
	3	3			2	3			5	0		
nivel	4	2		peso	4	3		ocup	2	5		

- Al no especificar alguna opción, se imprimen las estadísticas por default en las que se señala las tablas construidas (en este caso fueron 3 : edad vs nivel, ocup vs peso y edad vs ocup).

- Señala que datos fueron omitidos por estar fuera del rango marcado.

- Número de casos leídos.

- Para cada variable imprime algunas estadísticas descriptivas: media, desviación estandar, etc.

- Imprime las tablas.

- Bajo la hipótesis de independencia entre renglones y columnas, imprime la X^2 de Pearson y la X^2 de Yates y los valores esperados para cada tabla, así como los datos excluidos.

/DELTA

Cuando las frecuencias de las celdas son muy pequeñas, algunos investigadores prefieren añadir una constante a cada frecuencia de la celda (por lo general 0.5), esto es posible con la construcción:

/DELTA = (valor a ser añadido).

Opciones para la obtención de resultados:

/PRINT OBS.
Imprime las tablas de frecuencias a menos que se le especifique NO OBSERVED.

EXC.
Imprime la tabla con los casos de frecuencias excluídas a menos que se le especifique NO EXCLUDED.

LIST = (#).
Imprime sólo los casos que son excluídos de una o más tablas.

PERCENT = NO, ROW, COL, TOT.
Calcula los porcentajes por renglones (ROW), columnas (COL) o del total de la tabla (TOT).

Estas opciones por default ya están incluídas, sólo se indican en el programa si no se desean obtener.

Ejm Si no se quiere imprimir la tabla con la frecuencia de los valores observados :

/PRINT NO OBSERVED.

/STATISTICS
Se pueden obtener estadísticas, medidas de asociación, correlación, pruebas de hipótesis y otras estadísticas.

Opciones:

CHISQUARE. LRCHI.

FISHER. CORRELATION.

SPERMAN. (sólo 2x2) LAMBDA.

TAUS. MCNEMAR. (sólo rxr)

LINEAR. (sólo 2xc ó rx2)

CONTINGENCY.

TETRACHRORIC. (sólo 2x2)

GAMMA.

UNCERTAINY.

Ejm para obtener la X^2 y la Gamma:

```
/STATISTICS CHISQUARE.
            GAMMA.
```

Por otra parte, cuando se obtienen valores esperados muy pequeños, a veces es preferible 'colapsar' categorías para formar nuevas con valores esperados considerables. esto se puede hacer de la siguiente forma:

```
/STATISTICS    MINIMUM = 1.
/PRINT         EXPECTED.
```

(si el mínimo valor esperado es menor que 1 en una categoría, se 'colapsa' con la categoría adyacente)

Si se efectua el 'colapso' de categorías, las estadísticas **Ø**, el Coeficiente de Contingencia y la U de Cramer son calculadas con la X^2 de la tabla 'colapsada', todas las demás estadísticas son calculadas de la tabla original.

/PRINT

También es posible obtener bajo la hipótesis de independencia, los valores esperados, estandarizados y las desviaciones de Freeman-Tukey, las diferencias de los valores observados y esperados con las siguientes opciones:

EXPected FREeman STANdarized DIFference

Ejm para obtener los valores estandarizados :

```
/PRINT    STAND.
```

7.4.2 Tablas de n-criterios

Para formar tablas de n-criterios existen varios procedimientos (todos utilizan en comando TABLE) y que a continuación se especifican.

(Es importante señalar que todos los procedimientos van precedidos por los comandos PROBLEM, INPUT, VARIABLE, CATEGORY indicados anteriormente).

Para indicar que variables se van a tabular se utiliza:

```
/TABLE    INDICES = INGRESO,OCUP, PLANTEL.          (1)
```

Indica que PLANTEL es la variable que cambia más lentamente los índices e INGRESO es la que cambia más rápido, esta instrucción es similar a:

```
/TABLE    CATVAR = INGRESO.
          CATVAR =  OCUP.                           (2)
          CATVAR = PLANTEL.
```

también a:

```
/TABLE    COLUMN =INGRESO.
          ROW    =  OCUP.                           (3)
          CATVAR = PLANTEL.
```

Estas tres formas son similares y se pueden usar indistintamente.

Existen tres procedimientos para leer los datos, éstos dependen de cómo se haga la declaración de las variables:

1' Las formas anteriormente vistas son para declarar las variables sin especificar que se va a trabajar con frecuencias ya dadas, ésta es una forma de leer los datos. Si por ejemplo se

declararon las variables de la siguiente manera:

```
/VARIABLE NAMES = PLANTEL,OCUP,INGRESO.
/CATEGORY CODES(1) = 1,2,3,4,5.
          NAMES(1) = acc,nau,ori.sur,val.
          CODES(2) = 1,2.
          NAMES(2) = patron,obrero.
          CODES(3) = 1,2.
          NAMES(3) = 'pri ing','reing.
/TABLE    CATVAR   = INGRESO.
          CATVAR   =  OCUP.
          CATVAR   = PLANTEL.
```

los datos, que deben estar al final del programa después del /END

serán registrados de la siguiente forma

```
2 1 1
3 1 2
2 1 2
3 1 1
1 1 1
5 2 1
2 1 1
5 2 2
4 1 1
3 2 2
1 1 1
4 2 1              (**)
5 1 2
2 1 2
3 1 2
5 2 2
2 2 1
3 1 2
4 2 2
4 1 1
  .
  .
  .
```

donde los datos por ejemplo del primer renglón (2 1 1) quieren decir es un patrón de primer ingreso al plantel naucalpan.

2' Por otra parte, si se tienen las frecuencias, bastará

con indicarlas junto con los índices correspondientes, con el comando COUNT dentro de TABLE.

Sintáxis: /TABLE COUNT = <var de las frecuencias>.

Por ejm *, si se tienen los siguientes datos, que son equivalentes a (ff) y que deben estar al final del programa:

```
( plantel ocup ing frec)
      1    1   1    19
      1    1   2    37
      1    2   1   820
      1    2   2  1567
      2    1   1    22
      2    1   2    30
      2    2   1   628
      2    2   2  1190
      3    1   1    21
      3    1   2    38
      3    2   1   855
      3    2   2  1809
      4    1   1    21
      4    1   2    38
      4    2   1   663
      4    2   2  1412
      5    1   1    33
      5    1   2    60
      5    1   1   901
      5    2   2  2312
```

Las instrucciones serían:

```
              .
              .
              .
/TABLE    INDICES = INGRESO, OCUP, PLANTEL.
          COUNT = FREC.
              .
              .
              .
/END
```

y de esta manera se obtiene la tabulación de los datos.

3' La otra forma de declarar la tabla y que es más práctica para no declarar los índices, sólo las frecuencias, es incluyendo el commando /INPUT TABLE que es una opción de /INPUT que anteriormente se especificó.

Sintáxis: /INPUT TABLE = <no. de niveles>.
donde <no. de niveles> indica los niveles de la tabla con los índices en orden de acuerdo a la rapidez con que cambian (el primero es el más lento y el último es el más rápido).

ejm :

```
/INPUT    VARIABLES = 3.
          TABLE = 5,3,2.
          FORMAT = FREE.
/VARIABLE ...
/CATEGORY...
/TABLE ...
/END
datos
```

y los datos deberán tener la siguiente secuencia :

```
1@ 37 820
1567 22 30
.
.
.
60 901 2312
```

De acuerdo al análisis que se esté efectuando, a veces se desea separar una tabla en varias de menor dimensión y efectuar el análisis por cada tabla.

/CONDITION
Esta instrucción genera 2 tablas diferentes. Separa las categorías de la variable indicada.

```
/TABLE    INDICES = PLANTEL,INGRESO.
          CONDITION = UCUP.
```

El resultado son dos tablas de la siguiente forma :

	patron			obrero	
PLANTEL/INGRESO	pri ing	reingreso	PLANTEL/INGRESO	pri ing	reingreso
azc	18	37		820	1567
nau	22	30		628	1190
ori	21	38		855	1809
sur	21	38		663	1412
val	33	60		901	2312

3.5 Ajuste del modelo

BMDP cuenta con varias opciones (que a continuación se señalan) para el ajuste del modelo, el uso de cada una depende de las necesidades del investigador.

Cabe señalar que cuando se requieren agrupar 2 o más variables para indicar asociación entre ellas se debe utilizar la primera letra de la variable, seguida por un punto. Por ejm * la variable OI (OCUP e INGRESO) indica las siguientes cuatro categorías :

patron y primer ingreso
patron y reingreso
obrero y primer ingreso
obrero y reingreso

/FIT

Esta instrucción se utiliza para ajustar el modelo la primera letra de la variable será usada para representar el índice.

ASSOCIATION
Señala el grado de asociación.

Sintáxis: /FIT ASSOCIATION = (#).
donde (#) indica el grado de asociación.

Ejm si se desea obtener la G^2 y la X^2 de Pearson con un grado de asociación 2, se deberá teclear:

```
/PROBLEM  TITLE = 'PRUEBA'.
/INPUT    VARIABLES = 3.
          TABLE = 5,2,2.
/VARIABLE NAME = PLANTEL, OCUP, INGRESO.
/CATEGORY CODES(1) = 0,1,2,3,4,5.                    (A)
          NAMES(1) = azc,nau,ori,sur,val.
          CODES(2) = 0,1.
          NAMES(2) = patron,obrero.
          CODES(3) = 0,1.
          NAMES(3) = 'pri ing','reing'.
/TABLE    CATVAR   = INGRESO.
          CATVAR   =  OCUP.
          CATVAR   = PLANTEL.
/FIT      ASSOCIATION = 2.
/END.
```

(se obtendrá la G^2 y X^2 de Pearson para las asociaciones parciales y marginales de orden 0,1).

/FIT ALL

Se utiliza para ajustar todos los modelos jerárquicos de una tabla de 2 ó 3 criterios. En caso de tener una tabla de 4 criterios se podrían ajustar con esta instrucción n-tablas de 3 criterios (también si se tiene una tabla de 3 criterios se pueden ajustar n-tablas de 2 criterios) incluyendo la instrucción CONDITION señalada anteriormente.

Ejm * usando las intrucciones referidas en (A) :

```
.
. (A)
.
/FIT ALL.
/END.
```

Se obtendrá como resultado un cuadro con todos los modelos

requeridos con las estadísticas X^2 y G^2 de Pearson, los grados de libertad de ambos y el número de iteraciones efectuadas por cada modelo.

/FIT MODEL

Si se quiere ajustar un modelo específico se utiliza esta instrucción (como se señaló anteriormente la primera letra de la variable es la que se usa para identificarla).

Sintáxis: /FIT MODEL = (modelo).

Ejm:

1. Si queremos ajustar el modelo PLANTEL + OCUP + INGRESO + PLANTEL.INGRESO se tendrá lo siguiente:

   /FIT MODEL = P, O, I, PI.

2. BMDP trabaja con modelos jerárquicos, entonces si se tiene la instrucción

   /FIT MODEL = PI,PO,PI.

   se refiere al modelo

   PLANTEL + OCUP + INGRESO + PLANTEL.INGRESO + PLANTEL.OCUP + OCUP.INGRESO.

3. Si se quiere hacer el ajuste de dos modelos, bastará con poner:

   /FIT MODEL = PI,PO,OI.
   O,PI.

Como resultado de esta instrucción, se obtiene la X^2 y la G^2 con sus respectivos grados de libertad y el número de interacciones.

/FIT ADD y /FIT DELETE

Permite ajustar los modelos con el procedimiento 'STEPWISE'(añadir o quitar términos).

Para efectuar el ajuste añadiendo términos se utiliza /FIT ADD, las opciones para ésta son SIMPLE o MULTIPLE. El modelo inicial para el ajuste, es el especificado en /MODEL añadiendo el término (simple o múltiple) en turno e imprime el resultado de la prueba de ajuste del modelo y la diferencia de éste y el modelo original; después de ajustar todos los nuevos modelos, escoge al 'mejor'. Si en la instrucción STEP se le marca un número mayor que 1, entonces se reemplaza el modelo original por el 'mejor' modelo y otra vez añade términos. Esta operación se repite hasta el número de pasos especificado o cuando la prueba de 'ajuste del mejor modelo' y la prueba de la diferencia son no significativos (sus probabilidades son mayores que el criterio indicado en PROBABILITY).

Para efectuar el ajuste quitando términos se utiliza /FIT DELETE (con opciones SIMPLE o MULTIPLE) y el procedimiento es análogo al descrito anteriormente con la instrucción ADD, a diferencia de que el criterio para terminar el proceso es que se cumplan el número de pasos especificado en STEP o cuando la prueba de ajuste del 'mejor' modelo y de la diferencia sea significativa.

Para hacer el ajuste con ADD, se recomienda la opción MULTIPLE ya que de esta manera los términos con efectos

de orden mayor no pueden ser incluídos en una opción posterior, sin que todos los términos con efectos de orden menor incluídos en los de orden mayor sean significativos. Con DELETE, se recomienda la opción SIMPLE ya que con MULTIPLE se borran todos los términos de segundo orden y todos los involucrados de orden mayor, y visto de esta manera es muy extremoso.

3.6 Obtención y resultados del ajuste

/PRINT

En BMDP se pueden obtener otros indicadores en el ajuste, esto es posible con el comando /PRINT, después de haber indicado el modelo.

Sintáxis:

/PRINT	EXPected.	(imprime los valores esperados)
	STANdarized.	(imprime las desviaciones estandarizadas)
	FReeman.	(imprime las desviaciones de Freeman-Tukey)
	DIFFerence.	(imprime las diferencias entre los valores esperados y los observados)
	CHISQuare.	(imprime el valor de la X^2)
	LRCHI.	(imprime el valor de la G^2)

Ejm

```
/MODEL   P, I, O, P1, PO.
/PRINT   STAND.
         DIFF.
```

(del modelo especificado imprimirá las desviaciones estandarizadas y las diferencias entre los valores observados y esperados)

/PRINT

Es posible obtener también en caso de que la prueba sea no significativa, las estimaciones de los parámetros de modelos loglineales.

Sintáxis: /PRINT LAMBDA. (imprime los parámetros estimados y los estimados divididos entre su error estandar)

BETA. (imprime los parámetros multiplicativos $\beta = e^{\lambda}$)

VARIance. (imprime las matrices de correlación y covarianza entre los estimadores de los parámetros).

/TABLE EMPTY

Con este comando se pueden definir los 'ceros estructurales'.

Sintáxis: /TABLE EMPTY = <lista de indices>.
donde <lista de indices> indica los indices de las celdas que tienen 'ceros estructurales'.

Como resultado se obtiene una tabla con '1' y '0' en donde se marcó el 'cero estructural', también se obtienen los valores esperados y las desviaciones estandarizadas, ambos tomando en cuenta el 'cero estructural'.

/FIT CELL

Se pueden identificar las celdas con valores extremos utilizando esta instrucción. Si el resultado del ajuste del

modelo es no significativo, no se realiza el proceso.

Sintáxis: /FIT CELL = NO, STAN, FR.

Identifica las celdas con valores extremos (calculados paso a paso con la instrucción STEP usando los valores estandarizados o las desviaciones de Freeman-Tukey.

STEP = <#>.
<#> es el número máximo de celdas a identificar con valores extremos.

PROB = <#>.
<#> es el nivel de significancia.

```
Ejm #
/FIT  MODEL = F , O , I , FI.
      CELL = STAN.
      STEP = 4.
```

/FIT STRATA

Elimina cada categoría en turno para cada índice especificado en <lista>. Los niveles que tienen dos categorías no son eliminados.

Sintáxis: /FIT STRATA = <lista>.
ó /FIT STRATA = ALL.

```
Ejm       /FIT  MODEL  = F , O , I , FI.
                STRATA = ALL.
```

3.7 Facilidades del paquete BMDP

Las facilidades que ofrece BMDP son múltiples, a continuación se muestran algunas de éstas.

/STACK

Para imprimir una tabla en forma horizontal se utiliza esta instrucción.

Sintáxis: /TABLE STACK = <lista var>.

ejm * para imprimir en forma horizontal:

```
/TABLE     INDICES =  INGRESO, OCUP, PLANTEL.
           STACK   =  INGRESO, OCUP.
                      (variables que se van a imprimir en
                       forma horizontal).
```

Se obtendrá como resultado la siguiente tabla:

	OCUP			
	patron		obrero	
	INGRESO			
	pri ing	reing	pri ing	reing
PLANTEL				
azc	18	37	829	1567
nau	22	30	628	1190
ori	21	38	855	1809
sur	21	38	643	1412
val	33	60	901	2312

/PRINT MARG

Con esta instrucción se pueden imprimir los marginales de las tablas para los criterios de clasificación que se lo indique.

Ejm * para obtener los marginales de un criterio de clasificación:

```
/TABLE   INDICES =  INGRESO, OCUP, PLANTEL.
/PRINT   MARG    =  1.
```

imprimirá cada tabla con 1 criterio de clasificación los marginales:

OCUP

patron	obrero	total
318	6077	6506

igual para PLANTEL e INGRESO.

/PRINT PERCENT

Al igual que en tablas de 2x2 se pueden obtener los porcentajes, de la misma manera se pueden obtener en tablas con n-criterios.

Sintáxis. /PRINT PERCENT = RO, REN, COL, TOT.

/INPUT FILE y /SAVE FILE

A veces, cuando se tiene una gran cantidad de datos, resulta muy tedioso el que éstos se encuentren dentro del mismo programa que contiene las instrucciones de tabulación y ajuste del modelo. BMDP ofrece la facilidad de separar lo que es propiamente la declaración de los datos y el ajuste en sí. Esto de alguna manera permite que la lectura de los datos sea más rápida. También es muy conveniente esta forma de efectuar el ajuste, cuando se requieren hacer varias ocasiones un análisis o hacer transformaciones en los datos. La primera parte consiste en declarar los datos y las variables, formando un archivo con esta información con /SAVE FILE.

Sintáxis: /SAVE CODE = <nombre>.
FILE = <nombre de archivo>.
NEW.

donde CODE = <nombre> es usado para nombrar un archivo en BMDP. El nombre

asignado a CODE debe ser menor o igual a 8 caracteres. Deberá especificarse cada vez que el archivo BMDP se use como entrada de datos.
'nombre de archivo' es el nombre del archivo donde se guarda la información. .DAT.

La segunda parte, consiste en declarar el ajuste del modelo. Para acceder los datos que se salvaron en /SAVE FILE se utiliza /INPUT FILE.

Sintáxis: /INPUT CODE = 'nombre'.
FILE = 'nombre de archivo'.

donde CODE y FILE son los mismos que se declararon en /SAVE.

Ejm 1 Si se desea efectuar el ajuste, el primer archivo que contiene la declaración de los datos y variables, supónase que se llama PRUEB1.DAT, sería de la siguiente forma:

```
/PROBLEM  TITLE = 'declaracion de datos'.
/INPUT    VARIABLES = 5.
          CASES = 20.
/VARIABLE NAME = PLANTEL, COMP, INGRESO.
          COUNT = FREQ.
/CATEGORY CODES(1) = 1,2,3,4,5.
          NAMES(1) = a,c,m,an,pr,,sup,val.
          CODES(2) = 1,2.
          NAMES(2) = 'urban, sorera.
          CODES(3) = 1,2.
          NAMES(3) = 'pri ing','reing .
/SAVE     CODE = YO.
          FILE = DATOS.
          NEW
/END

          1   1   1   18
          1   1   2   37
          1   2   1  829
          1   2   2 1567
          2   1   1   32
          2   1   2   50
          2   2   1  626
          2   2   2 1190
          3   1   1   21
          3   1   2   38
```

```
3   2   1   855
3   2   2  1809
4   1   1    21
4   1   2    38
4   2   1   663
4   2   2  1412
5   1   1    33
5   1   2    60
5   1   1   901
5   2   2  2312
```

Este archivo se ejecuta con el programa P1D de la siguiente forma:

```
$BMDF P1D PRUEBA.DAT  PRUEBA.SAL
```

El segundo archivo que contiene el ajuste del modelo sería de la siguiente forma:

```
/PROBLEM  TITLE = 'ajuste del modelo'.
/INPUT    FILE  = DATOS.
          CODE  = YO.
/TABLE    INDICES = INGRESO, OCUP, PLANTEL.
/FIT ALL.
/END
```

Este archivo se ejecuta con el programa P4F, señalado anteriormente. Al ejecutar este programa, se efectúa el ajuste que se haya requerido.

/SAVE CONTENT e /INPUT CONTENT

En algunos casos, donde se procesa gran cantidad de datos es necesario guardar las tablas tabuladas (que se hayan obtenido en una ejecución del programa), en un archivo para posteriormente leerlos directamente sin tener que tabular nuevamente. Para esto se utliza esta instrucción, después de /TABLE. Con el nombre que se le asigne creará

las tablas y para identificarlas les pone un número consecutivo. También se deben incluir las instrucciones FILE y CODE que marcan en que nombre y código se van a guardar.

Sintáxis: /SAVE CONTENT = TABLE,DATA.
FILE = <nombre de archivo>.
CODE = <nombre>.

donde: CONTENT = TABLE, DATA. quiere decir qué tablas o datos se guardarán en el archivo BMDF.
FILE y CODE tienen las mismas especificaciones que la instrucción anterior.

Ejm # si se quieren generar y guardar 3 tablas marginales

```
      .
      .
      .
/TABLE    COLUMN = PLANTEL, INGRESO.
          ROW    = INGRESO, OCUP.
/TABLE    COLUMN = PLANTEL.
          ROW    = OCUP.
/SAVE     CONTENT = TABLE.
          FILE= TABLAS.
          CODE = YO.
          NEW
          .
          .
          .
```

generará las siguientes tablas llamadas TABLE1, TABLE2, TABLE3, en el archivo TABLAS con el código YO:

```
TABLE1     PLANTEL vs INGRESO
TABLE2     INGRESO vs OCUP
TABLE3     PLANTEL vs OCUP
```

y por cada tabla numera las variables. Por ejemplo, en la tabla TABLE1 asigna 1 a PLANTEL y 2 a INGRESO.

Para cada una obtiene la X^2 de Pearson y el valor mínimo esperado.

Para leer las tablas guardadas con la instrucción /SAVE

CONTENT se utiliza /INPUT CONTENT

Sintáxis: /INPUT CODE = <nombre>.
CONTENT= TABLE.DATA.
FILE = <nombre del archivo>.

CODE, CONTENT y FILE son las asignadas con el comando SAVE CONTENT.

Ejm. Para accesar la tabla TABLE3 (anteriormente señalada) :

```
/PROBLEM  TITLE = 'LEE LA TABLA TABLE3'.
/INPUT    CODE = YO.
          CONTENT = TABLE3.
          FILE = TABLAS.
/TABLE    COLUMN = 1.
          ROW = 2.
/FIT .....
/END.
```

con la instrucción /TABLE se indica la variable con el número que se le asignó con SAVE y forma la tabla como se le haya indicado.

/FIT CONVERGENCE y /FIT ITERATION

Se puede controlar el valor del criterio de convergencia con este comando, así como el número máximo de iteraciones con /FIT ITERATION.

Sintáxis: /FIT CONVERGENCE = <#1>, <#2>.

donde:
<#1> es la diferencia absoluta máxima permitida.
<#2> es la diferencia máxima entre los observados totales y marginales totales en el modelo.

Sintáxis: /FIT ITERATION = <#>.

donde:
<#> es el máximo de iteraciones para

ajustar cualquier modelo.

Ejm

```
.
.
.
/FIT MODEL = IO,PO.
     ITERATION = 10.
     CONVERGENCE = .001,.0001.
.
.
.
```

3.8 ESTRUCTURA GENERAL DE UN PROGRAMA EN BMDP

```
/PROBLEM   TITLE
/INPUT     VARIABLES
           CASES
           FORMAT
/VARIABLE  NAMES
/CATEGORY  CUTPOINTS
           CODES
           NAMES
/TABLE     INDICES
       ó
           COL
           ROW
           CAT

         * PAIR. ó CROSS.
         * CONDITION
         * COUNT
         * EMPTY
         * STACK
         * DELTA
* /PRINT   OBServed
           EXCluded
           LIST
           EXPected
           LAMBDA
           VARiance
           BETA
           STANDardized
           FREEman
           ADJust
           DIFFerence
           CHISQ
           MARGins
           PERCent
* /STATistics CHISQuare
           CONTingency
           LRCHI
           FISHer
           TETRachoric
           CORRelation
           SPEARman
           GAMma
           LAMBda
           TAUS
           UNCertainty
           MCNemar
           MINimum
           LINEAR
```

```
  /FIT        ALL
              SIMULtaneous
              ASSOCiation
              MODEL
            * ITERation
            * CONVergence
            * ADD
            * DELETE
            * STEP
            * PROBability
            * CELL
            * STRATA
* /INPUT      TABLE
              CONTent
* /SAVE       CONTent
```

* Opcionales

Nota : Las letras mayúsculas indican los nombres que BMDP reconoce, es decir, que indicando éstos es suficiente para que BMDP los tome en cuenta y de esta manera, no escribir el nombre completo.

Tabla comparativa de algunos resultados que se obtienen con los paquetes GLIM, SYSTAT y BMDP.

	GLIM	SYSTAT	BMDP
Valores esperados	*	*	*
Devianza	*	*	*
χ^2	*	*	*
Estimación de parámetros	*		*
Residuales (obs-esp / $\sqrt{esp}$)	*	*	*
Desviaciones de Freeman-Tukey			*
Diferencias obs-esp		*	*
Porcentajes (columnas y renglones)	*	*	*
Coeficiente de contingencia			*
Probabilidades exactas de Fisher			*
Correlación de Spearman			*
Medidas de Goodman-Kruskall			*
Métodos para la selección del "mejor" modelo			*
Cero estructural		*	*
Colapsar categorías			*
Opción para añadir ó = .x		*	*
Criterio para terminar las iteraciones	*	*	*
Comando de "ayuda" incluído en el paquete		*	*
Lectura de datos en archivos externos	*	*	*

Tiempo de respuesta para el ajuste de diferentes modelos

celdas	20		40			100			
	10x2	10x4		5x2x4	10x10		5x2x10		5x2x2x5
	3. 4'30"	*3.* 56'	*1.* 10'		*3.* + 7hrs	*1.* 1hr40"		*6.* 4hrs30'	
GLIM	*4.* 48"	*4.* 1'	*2.* 1'30"		*4.* 5'	*2.* 05"		*5.* 05'	
	3. 04"	*3.* 05"	*1.* 08"		*3.* 10"	*1.* 25"		*6.* 48"	
SYSTAT	*4.* 04"	*4.* 05"	*2.* 06"		*4.* 10"	*2.* 15"		*5.* 30"	
	3. 14"	*3.* 14"	*1.* 15"		*3.* 15"	*1.* 16"		*6.* 16"	
BMDP	*4.* 14"	*4.* 14"	*2.* 14"		*4.* 15"	*2.* 16"		*5.* 17"	

celdas	500			1000		
	10x50	10x5x10		10x100	10x10x10	
	3. chec alloc	*1.* chec alloc		*3.* chec alloc	*1.* chec alloc	
GLIM	*4.* chec alloc	*2.* si ???		*4.* chec alloc	*2.* 2hrs	
	3. out	*1.* 1'30"		*3.* out	*1.* out	
SYSTAT	*4.* out	*2.* 1'03"		*4.* out	*2.* 5'	
	3. 17"	*1.* 18"		*3.* 20"	*1.* 28"	
BMDP	4 18"	*2.* 17"		*4.* 22"	*2.* 25"	

celdas	1000	2000
	4x5x10x5	10x100x2
	6. chec alloc	chec alloc
GLIM	*5.* 2hrs 10"	
	6. 05'	out
SYSTAT	*5.* 02'	
	6. 30"	*1.* 33"
BMDP	*5.* 25"	*2.* 31"

Modelos

1. x1 + x2 + x3 + x12 + x23 + x13

2. x1 + x2 + x3

3. x1 + x2 + x12

4. X1 + X2
5. X1 + X2 + X3 + X4
6. X1 + X2 + X3 + X4 + X12 + X13 + X14 + X23 + X24 + X34 + X123 + X124 + X234 + X1234

Como se puede observar, en GLIM el tiempo de respuesta que tarda en efectuar el ajuste, depende del número de celdas y del modelo. Dado un número de celdas, el tiempo de respuesta para el ajuste de un modelo con menos términos es menor. Conforme aumenta la cantidad de celdas, el tiempo de respuesta aumenta considerablemente, por ejemplo para el modelo *1*. va de 10" para 40 celdas, hasta 1 hr 40" para 100 celdas.

Cuando se tiene el mismo número de celdas, en general, en GLIM el tiempo de respuesta es menor cuando se tienen más variables.

En BMDP, los tiempos de respuesta no varían mucho, ya que por ejemplo van de 15" con 40 celdas hasta 30" con 2000 celdas, para el modelo *1*. En comparación con SYSTAT y GLIM estos tiempos son mínimos.

En SYSTAT los tiempos de respuesta no varían demasiado y tan extremadamente como en GLIM. De acuerdo a éste se puede decir que están en un término medio respecto a GLIM y BMDP.

Por otra parte, en cuánto al límite de manejo de datos, se observa que con GLIM y SYSTAT en general, no se pueden manejar 1000 celdas y en éste sentido BMDP no tiene limitaciones (de acuerdo al manual, se pueden manejar hasta 5000 celdas).

C O N C L U S I O N E S

El uso de paquetes estadísticos debe considerarse como una herramienta de gran utilidad para el análisis de datos, sin hacer a un lado los aspectos teóricos que llevan a un buen análisis, ya que los diseñadores de paquetes no consideran el problema de la organización inicial de la información.

La utilización de esta herramienta permite al investigador no distraer su atención en cálculos numéricos que mediante el uso de paquetes estadísticos son relativamente rápidos y precisos.

De los 3 paquetes estadísticos que fueron presentados para el análisis de Tablas de Contingencia, no se puede recomendar uno en especial o decir que alguno es el mejor en todos los sentidos, ya que cada uno tiene ventajas y desventajas y la elección de que paquete utilizar depende de las necesidades que el investigador requiera.

Las ventajas que ofrece el paquete GLIM son las siguientes: permite el cálculo de los parámetros (u's), maneja archivos que facilitan el uso de este paquete evitando de alguna manera repetir procesos, está disponible para una microcomputadora PC . Las desventajas que se deben considerar son: tiene límites en cuánto a la capacidad de almacenar información (sólo se pueden accesar aproximadamente 500 celdas), no contiene 'help' que a

veces facilita el uso del paquete y en cuánto al tiempo de respuesta se considera que es muy lento.

Por otra parte las ventajas del paquete SYSTAT son: el tiempo de respuesta es relativamente rápido, contiene un 'help' que permite que el paquete sea 'amigable', se puede accesar en una microcomputadora PC. Las desventajas son: no permite el cálculo de los parámetros (u's), tiene límites en cuánto a capacidad de almacenar información (sólo permite aproximadamente 500 celdas).

Finalmente, considerando que BMDP es un paquete estadístico muy completo, las ventajas que ofrece son considerables, entre éstas se encuentran: permite el cálculo de parámetros y otro tipo de estadísticas que los paquetes GLIM y SYSTAT no consideran, en cuánto al tiempo de respuesta es rápido, almacena gran cantidad de información, contiene 'help' que de alguna manera facilita su uso. La desventaja que tiene este paquete es que sólo se encuentra disponible en una computadora VAX, que por lo general no es fácil el acceso a este tipo de computadoras.

ANEXO: RELACION ENTRE LOS PARAMETROS OBTENIDOS EN LOS PAQUETES GLIM Y BMDP

Dados los detalles de los paquetes GLIM y BMDP, es importante establecer algunas relaciones entre los parámetros (u's) que obtienen estos paquetes estadísticos.

Analizando los parámetros obtenidos por GLIM se observa que para el cálculo de dichos parámetros no sigue las relaciones establecidas señalada en los capítulos anteriores.

Por otra parte, BMDP si se basa en estas relaciones para el cálculo de los parámetros.

Sin embargo, se pueden obtener ciertas correspondencias entre estas dos formas de calcular los parámetros.

En base a lo anteriormente expuesto, para dos criterios se tiene que bajo Ho:Pi.P.j

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}$$

Los parámetros obtenidos son (2.1.10),(2.1.11), (2.1.12), de la primera parte de esta Tesis. La demostración de esta igualdad consiste únicamente en sustituir los parámetros u, $u_{1(i)}$, $u_{2(j)}$ y simplificar (tomando en cuenta la hipótesis de independencia).

Para 3 criterios bajo la hipótesis de independencia, los parámetros obtenidos son (2.3.4) - (2.3.7) expuestos en la primera parte de esta Tesis. En caso de no haber independencia entre

las tres variables, los parámetros de asociación entre pares de variables se pueden calcular de la siguiente manera:

$$u_{12(ij)} = \frac{1}{K}\sum_{k=1}^{k} \log m_{ijk} - u_{1(i)} - u_{2(j)} - u$$

de una forma similar se calculan los parámetros $u_{13(ij)}$ y $u_{23(ij)}$.

De esta manera son calculados los parámetros en BMDP.

Por otra parte, haciendo un análisis y observando los resultados del cálculo de los parámetros obtenidos por GLIM se tiene que para dos criterios, bajo la hipótesis de completa independencia:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}$$

con

$$u = \log m_{11}$$

$$u_{1(i)} = \log m_{i1} - u$$

$$u_{2(j)} = \log m_{1j} - u$$

que de la misma manera anterior este resultado se demuestra simplemente sustituyendo valores y simplificando, tomando en cuenta la hipótesis de independencia $\log m_{ij} = m_{i.} m_{.j}$.

En particular para $i = 1$ y $j = 1$ se tiene:

$$u_{1(1)} = \log m_{11} - u = \log m_{11} - \log m_{11} = 0$$

$$u_{2(1)} = \log m_{11} - u = \log m_{11} - \log m_{11} = 0$$

Como se puede observar en GLIM, en el caso de Tablas de Contingencia de dos criterios, bajo la hipótesis de completa independencia, se "fijan" dos parámetros iguales a cero $u_{1(1)}$ y $u_{2(1)}$.

Siguiendo este método, para tres criterios se tiene que bajo la hipótesis de completa independencia:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

con

$$\begin{aligned} u &= \log m_{111} \\ u_{1(i)} &= \log m_{i11} - u \\ u_{2(j)} &= \log m_{1j1} - u \\ u_{3(k)} &= \log m_{11k} - u \end{aligned} \qquad (A1)$$

En particular para i = 1, j = 1, k = 1, se obtendrá que:

$$u_{1(1)} = u_{2(1)} = u_{3(1)} = 0$$

Como se puede ver, para tres criterios, bajo la hipótesis de completa independencia se "fijan" tres parámetros iguales a cero.

En el caso de tener asociaciones entre las variables, los parámetros se obtendrían de manera similar, con una pequeña variación. P.e. para obtener el parámetro u_{12} en el caso de una Tabla de Contingencia de IxJxK, se calcula de la siguiente forma:

$$u_{12(ij)} = \log m_{ij1} - u_{1(i)} - u_{2(j)} - u$$

con u, u_1, u_2 como en (A1).

B I B L I O G R A F I A

AGRESTI, A. 1984. "Analysis of Ordinal Categorical Data". John Wiley & Sons, New York.

BAKER R.J. AND NELDER J.A. 1978. "Generalized Linear Interactive modelling. Manual". The GLIM SYSTEM, Release 3. Numerical Algorithms Group, Oxford.

BISHOP, Y.M.M. FIENBERG S.E. AND HOLLAND, P.W. 1975. "Discrete Multivariate Analysis: Theory and Practice". MIT Press, Cambridge, Massachusetts.

BROWN M. B., L. ENGELMAN, J.W. FRANE, M.A. HILL, R.I. JENRICH, J.D. TOPOREK. 1983. "BMFL Statistical Software". University of California Press.

EVERITT, B. S. 1977. "The Analysis of Contingency Tables". London Chapman and Hall.

FIENBERG, S.E. 1979. "The Analysis of Cross-Classified Categorical Data". MIT Press, Cambridge.

LELAND WILKINSON, EVANSLON, IL. 1985. "SYSTAT. The System for Statistics". SYSTAT, Inc.

NELDER, J. A. 1974. "Loglinear Models for Contingency Tables: A Generalization of Classical Least Squares", Appl. Statist. 13, No. 3, pp. 323-329.

NELDER, J. A. AND WEDERBURN, R. W. M. 1972. "Generalized Linear Models". J. Roy. Statist. Soc. A 135, pp.370-384.