ESPECIALIDAD EN METODOS ESTADISTICOS

PRUEBAS DE INDEPENDENCIA CONDICIONAL PARA VARIABLES DISCRETAS (DESCRIPCION Y APLICACIONES)

Trabajo recepcional que como requisito parcial para obtener el diploma de esta Especialidad presenta:

JULIA AURORA MONTANO RIVAS

TUTOR: DR. Manuel Martínez Morales

XALAPA, VER. DICIEMBRE DE 1994

DATOS DEL AUTOR: Julia Aurora Montano Rivas, nació en Xalapa, Veracruz, en 1964. Realizó estudios primarios y secundarios en La Estanzuela, Ver, después se transladó a la ciudad de Xalapa donde realizó el bachillerato. En septiembre de 1983 ingresó a la Facultad de Estadística de la Universidad Veracruzana. Obtuvo el título de Licenciado en Estadística en 1990, con tesis intitulada "METODOLOGIA DE SUPERFICIE DE RESPUESTA". Desde 1988 a la fecha ha colaborado en el Laboratorio de Investigación y Asesoría Estadística (LINAE), en la Facultad de Estadística de la Universidad Veracruzana, como coordinadora de asesorías y responsable de los recursos bibliográficos y materiales didácticos con los que cuenta el LINAE. Además ha colaborado en proyectos de investigación con el Dr. Mario Miguel Ojeda. Paralelo a esto, en 1988 ingresó como docente en la Facultad de Estadística. En diciembre de 1994 concluyó sus estudios de Especialidad en Métodos Estadísticos.

AGRADECIMIENTOS:

GRACIAS SEÑOR DIOS por iluminarme mi mente.

Expreso mi más sincero agradecimiento

Al Dr. Manuel Martínez M. por todo su apoyo, comprensión y paciencia para la realización de este trabajo.

Al Dr. M.Miguel Ojeda R. por su enseñanza, amistad y apoyo moral para la realización de la Especialidad.

A mis Padres y a mi Hermana por todo su apoyo moral, espiritual y económico.

El Comité Académico de la Especialidad en Métodos Estadísticos, y el tutor del trabajo recepcional, autorizan la impresión y constitución de tribunales para la defensa.

Coordinador de la Especialidad
DR. Mario Miguel Ojeda Ramírez

Vocal
Lic. Víctor Manuel Méndez Sánchez

Director de la Facultad
Lic. Claudio Rafael Castro López.

Vocal
Lic. Sergio Hernández González

Tutor

INDICE

1. INTRODUCCION.

Los conceptos de independencia e independencia condicional son fundamentales en la teoría de inferencia estadística, así como los términos de suficiencia y ancilaridad, los cuales pueden derivarse del concepto básico de independencia condicional (Dawid, 1980).

Los diagramas de influencia probabilista permiten representar relaciones de dependencia/independencia condicional y permiten describir graficamente ciertas interrelaciones entre variables aleatorias. En este contexto nos ocuparemos del análisis de éstos, ya que son de mucha utilidad en problemas de estadistica aplicada.

En este trabajo se presenta resumidamente la metodología propuesta por Kullback (1968) para probar independencia e independencia condicional para el caso de variables discretas; pruebas que se basan en medidas de informacón (entropía). En seguida se describe la forma en que se emplean los diagramas de influencia probabilísta (Balow y Braganca, 1990) para representar relaciones de dependencia/independencia en un conjunto de variables.

En las secciones 2 y 3 se dan las definiciones usuales de independencia e independencia condicional. La sección 4 contiene una descripción de los diagramas de influencia probabilísta y dos ejemplos de su aplicación. Las pruebas de independencia e independencia condicional propuestas por Kullback se presentan en la sección 5.

Finalmente en la última sección (6) se presentan dos aplicaciones de la metodología descrita. La primera de ellas analiza la relación de algunas variables explicativas (sexo, edad, diabetes, localización del infarto) con la mortalidad en pacientes afectados por infarto agudo al miocardio. En la segunda aplicación se estudia la asociación entre un conjunto de variables (edad, sexo, nivel socioeconómico, ocupación) y la preferencia electoral según una encuesta levantada previamente (julio 1994) a la elección presidencial.

Para el lector interesado en profundizar en la aplicación de estas pruebas se proporciona al final la bibliografía referente al tema.

2. INDEPENDENCIA PROBABILISTA.

Sean X y Y variables aleatorias discretas para las cuales denotamos por:

p(x,y) la función de probabilidad conjunta de X, Y;
p(x) la función de probabilidad marginal de X;
p(x/y) la función de probabilidad condicional de X dado Y=y, definida por

$$p(x/y) = \frac{p(x,y)}{p(y)}$$

donde $p(y) \neq 0$

Definición. Sean X y Y variables aleatorias discretas, se dice que X y Y son estocásticamente independientes si

$$p(x,y) = p(x)p(y);$$

o equivalentemente, si

$$p(x/y) = p(x)$$

Si X y Y son independientes lo denotamos por $\mathbf{X \amalg Y}$.

3. INDEPENDENCIA CONDICIONAL.

Definición. Si se tienen X, Y, Z variables aleatorias, decimos que Y es condicionalmente independiente de X dado Z=z, si se cumple que para todo z:

$$p(x, y / z) = p(x / z)p(y / z) \qquad (1)$$

o equivalente si

$$p(x / z, y) = p(x / z) \qquad (2)$$

Notación: $X \coprod Y / Z$ denota que X es condicionalmente independiente de Y dado Z.

4. DIAGRAMAS DE INFLUENCIA.

Las interrelaciones de cierto tipo entre variables aleatorias pueden representarse mediante diagramas de influencia probabilista también conocido como redes probabilistas (Sucar, 1993) que son gráficas acíclicas dirigidas.

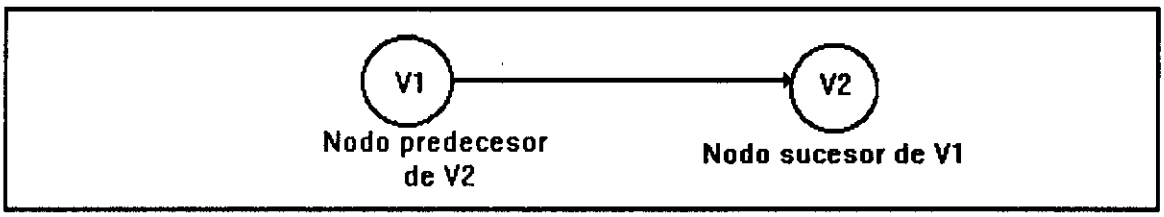
Principalmente pueden representarse relaciones de dependencia asimétricas; es decir en problemas en los cuales las variables de interés pueden separarse en forma natural en variables dependientes y variables independientes o explicativas;. es posible emplear diagramas de influencia probabilísta para representar esas relaciones.

A cada nodo de estos diagramas corresponde una variable aleatoria y tiene asociada la función de probabilidad condicional de la variable del nodo dadas las variables en los nodos antecesores. Puede demostrarse (Braganca, 1990) que a una gráfica de este tipo le corresponde unívocamente una función de densidad conjunta para todas las variables representadas.

Una de las características de los diagramas de influencia probabilistas es que debe de tener al menos un nodo inicial (sin antecesores)y un nodo terminal (sin sucesores).

Si dos variables son condicionalmente independientes dadas algunas de las otras variables entonces no habrá arco alguno uniendo los nodos correspondientes (Barlow y Braganca, 1990).

Es muy importante la dirección del arco ya que por medio de esta sabemos cuales son los nodos predecesores o sucesores decada nodo. Esta dirección se denota $[v_1, v_2]\cdot$; es decir, representa la relación:



Hay dos problemas relacionados con la aplicación de estos diagramas y los problemas de independencia condicional.

1) Determinar la topología del diagrama que implica la aplicación sistemática de sucesivas pruebas de independencia condicional.

2) Estimar las probabilidades condicionales y emplearlas posiblemente, (sobre la topología encontrada en (1)) con fines predictivos.

En la literatura, para el caso de variables discretas se proponen dos maneras de probar independencia condicional: (a) a través de modelos log-lineales (Agresti, 1984), (b) aplicando las pruebas propuestas por S. Kullback.

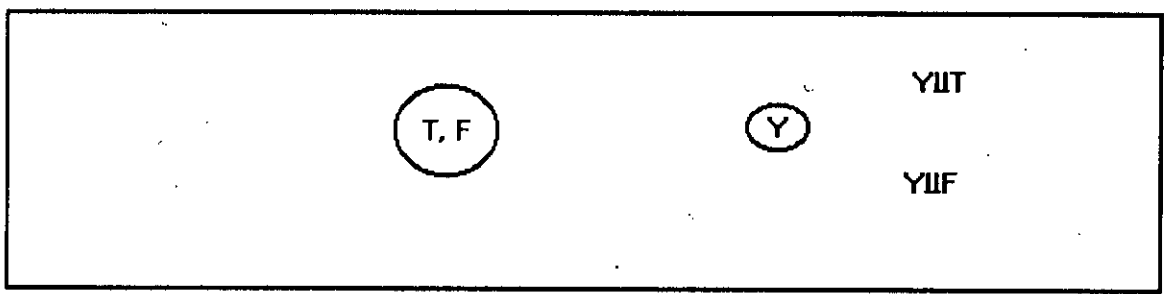
4.1 Ejemplo I. (Martínez M., Montano A.,Ojeda M., 1992)

Se realizó un experimento en los viveros de Xalapa para determinar cual combinación de sustratos y fórmulas de fertilización producía el mayor porcentaje de crisantemos que cumplieran con la norma de calidad exigida por ciertos compradores. La variable asociada con la calidad era el diámetro de la flor; considerándose de calidad aceptable si $d \geq 15$ cms. y no aceptable en caso contrario. Sea Y la variable definida por:

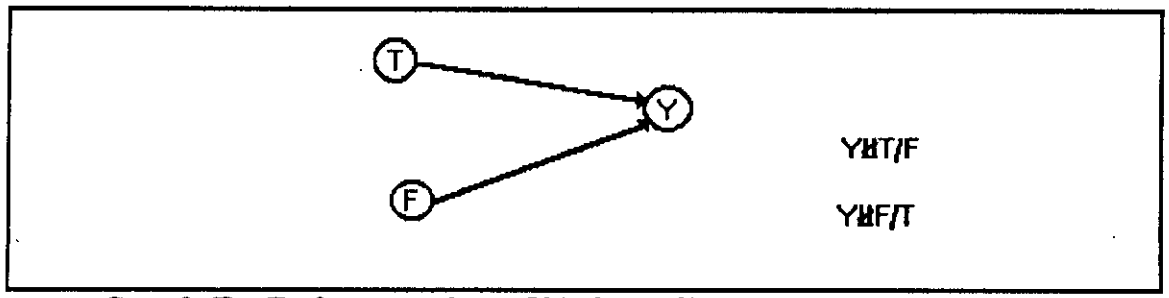
$$Y = \begin{cases} 1 \text{ si } d \geq 15 \\ 0 \text{ si } d < 15 \end{cases}$$

Se eligieron cuatro distintos tipos de sustratos (T=1,2,3,4), y dos fórmulas de fertilización (F=1,2).

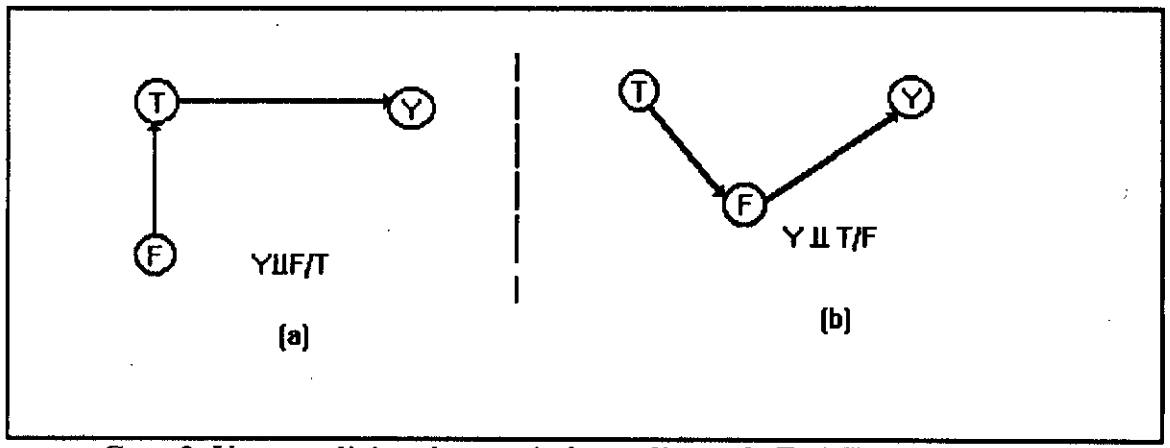
Se desea saber qué efectos tienen las variables T y F sobre Y. En términos de diagramas de influencia probabilistica se tienen varias probabilidades.



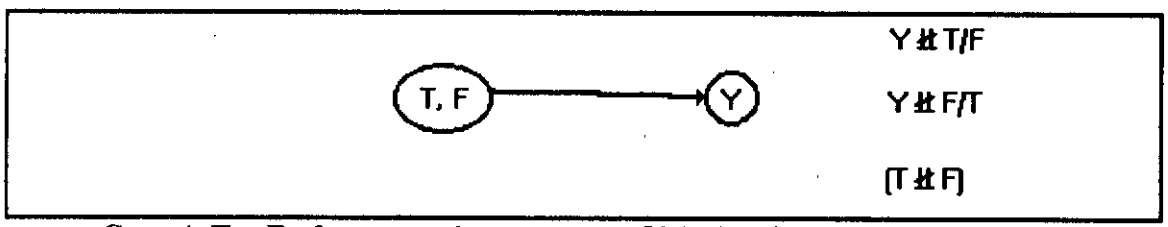
Caso 1. Las variables T y F no tiene efecto alguno sobre la variable Y (Y es independiente de T y F).



Caso 2. T y F afectan ambas a Y independientemente (no existe interacción).



Caso 3. Y es condicionalmente independiente de F (ó T) dado T (ó F).



Caso 4. T y F afectan conjuntamente a Y (existe interacción, no se pueden separar los efectos de T y F)

En el estudio original de este problema (Montano, A., Ojeda, M., 1991) se empleó el análisis de varianza convencional del cual se concluyó que el mejor tratamiento era la combinación F=1 y T=1.

Discretizado el problema mediante el empleo de la variable Y realizamos un análisis cualitativo basado en las distribuciones condicionales estimadas de Y/(T,F); Y,F/T y Y,T/F; ya que comparando estas distribuciones podrá inferirse si Y es condicionalmente independiente de T dado F o viceversa.

A continuación se muestran los resultados obtenidos en el experimento.

F=1

Y	1	2	3	4	TOTAL
0	1	3	2	1	7
1	3	1	2	3	9
TOTAL	4	4	4	4	16

F=2

Y	1	2	3	4	TOTAL
0	3	3	3	3	12
1	1	1	1	1	4
TOTAL	4	4	4	4	16

TABLA 1. Frecuencias correspondientes a las combinaciones de los cuatro diferentes sustratos (T) con los dos tipos de fertilización (F) para los dos niveles de calidad (Y) aceptables en el mercado.

Puede observarse que para F=2, el efecto de T es nulo; esto es, se observa la misma respuesta para T=1,2,3,4, dado F=2. Cuando F=1, T si parece tener un efecto diferenciador. Los datos sugieren probar la hipótesis:

Ho: Y ⫫ T/F VS Ha: Y ⫫̸ T/F

Kullback (1968) proporciona una prueba para hipótesis de independencia condicional como ésta para tablas de contingencia. El estadístico de prueba tiene una distribución (aproximadamente) Chi-cuadrada. Para nuestro ejemplo el valor del estadístico fue de $\chi^2 = 2.89$ con 6 grados de libertad. El resultado es no significativo al nivel $\alpha = 0.05$, lo que nos indica no rechazar Ho, es decir la hipótesis de independencia condicional de Y y T dado F. Esto es que lo determinante es la fórmula de fertilización. En términos gráficos Ho se representa por el diagrama 3(b).

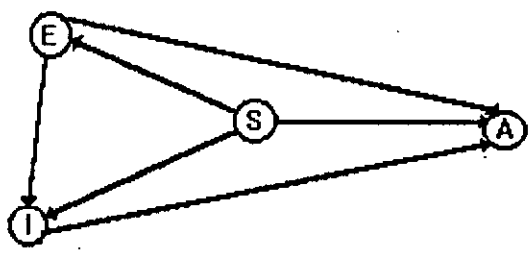
4.2 Ejemplo II. (Martínez M., Montano A., Ojeda M., 1992)

En una encuesta de opinión levantada en mayo de 1991 en la ciudad de Xalapa, interesaba, entre otras cosas, ver el efecto de algunas variables "independientes" sobre la opinión que se tenía sobre el aborto. Uno de los autores aplicó en aquella ocasión un modelo lineal generalizado para analizar los datos (Ojeda M. M., 1992). A continuación se describe la reformulación y análisis del problema empleado en diagramas de influencia probabilista y el concepto de independencia condicional.

Sean A, E, I, S las variables que representan la opinión sobre el aborto, escolaridad, ingreso mensual y sexo respectivamente.

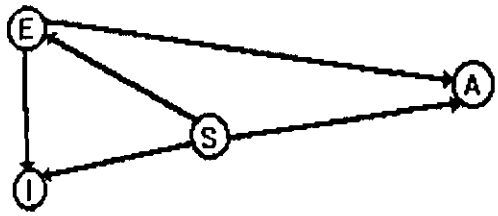
La hipótesis inicial es que las variables "independientes" (E,S,I) afectan a la variable A y existen también interrelaciones entre ellas.

El diagrama (hipótetico) inicial sería pues el siguiente:



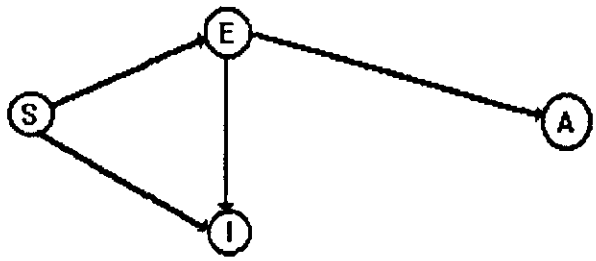
Aplicando repetidamente la prueba de Kullback (v. g. para determinar si $A \perp\!\!\!\perp I / (E,S)$ se fué reduciendo al diagrama de la siguiente manera:

[1] $A \perp\!\!\!\perp I/E,S$



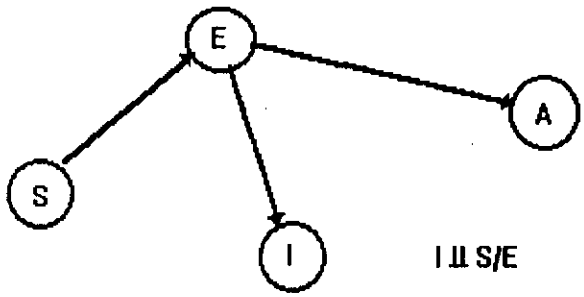
Esto es, desaparece el arco uniendo a los nodos I con A. (El ingreso no "afecta" la opinión sobre el aborto dadas la escolaridad y el sexo).

[2] $A \perp\!\!\!\perp I, S/E$



Dada la escolaridad, el sexo no "afecta" la opinión sobre el aborto. Es decir, dentro de cada nivel de escolaridad no hay diferencia entre las respuestas de los sexos. Adicionalmente se probó que I es condicionalmente independiente de S dado E:

[3]



La conclusión, clara en el diagrama, es que A es condicionalmente independiente de S e I, dado E. En otras palabras, si conocemos la escolaridad, el sexo y el ingreso no agregan información significativa acerca de la opinión sobre el aborto. En concreto, la gente con mayor escolaridad tiende a dar una opinión negativa sobre permitir el aborto deseado.

5. PRUEBA DE INDEPENDENCIA E INDEPENDENCIA CONDICIONAL.

Sea X la variable aleatoria discreta que toma valores en el conjunto $\{x_1, x_2, \ldots, x_r\}$ con probabilidades correspondientes $p(x_1), p(x_2), \ldots, p(x_r)$. Entonces la información (entropía) asociada con la distribución de X se define como:

$$H(X) = -\sum_{i=1}^{r} p(x_i) \ln p(x_i)$$

donde la base de los logaritmos es irrelevante. (De aquí en adelante usaremos logaritmos naturales).

Sea X la variable aleatoria discreta como se define anteriormente y sea Y una variable aleatoria discreta con valores $\{y_1, y_2, \ldots, y_c\}$ y probabilidades correspondientes a $p(y_1), p(y_2), \ldots, p(y_c)$. Entonces la información asociada con la distribución de X dado que $Y = y_j$ es, para cada j = 1,2,...,c :

$$H(X / Y = y_j) = -\sum_{j=1}^{c}\sum_{i=1}^{r} p(x_i / y_j) \ln p(x_i / y_j);$$

y la información media sobre X dada por Y se define entonces:

$$H(X / Y) = -\sum_{j=1}^{c} p(y_j) H(X / Y = y_j)$$

$$= -\sum_{j=1}^{c}\left[p(y_j) \sum_{i=1}^{r} p(x_i / y_j) \ln p(x_i / y_j) \right]$$

$$= -\sum_{j=1}^{c}\sum_{i=1}^{r} p(y_j) p(x_i / y_j) \ln p(x_i / y_j)$$

$$= -\sum_{j=1}^{c}\sum_{i=1}^{r} p(x_i, y_j) \ln p(x_i / y_j).$$

Se define la ganacia en información acerca de X dada por Y por

$$I(X / Y) = H(X) - H(X / Y).$$

Las pruebas de independencia e independencia condicional propuestas por Kullback se basan en la cantidad de información (o entropía) asociada con una variable aleatoria X y con la cantidad media de información proporcionada por Y acerca de X.

Como en realidad las funciones de probabilidad no se conocen tienen que estimarse de una muestra empleándose como estimadores de $p(x_i), p(x_i, y_j)$ y $p(x_i / y_j)$ respectivamente:

$$\hat{p}(x_i) = \frac{f_{i\bullet}}{N}$$

$$\hat{p}(y_j) = \frac{f_{\bullet j}}{N}$$

$$\hat{p}(x_i, y_j) = \frac{f_{ij}}{N}$$

$$\hat{p}(x_i / y_j) = \frac{f_{ij}}{f_{\bullet j}}$$

donde

$$f_{i\bullet} = \sum_{j=1}^{c} f_{ij}$$ para cada i = 1,2,...,r.

$f_{\bullet j} = \sum_{i=1}^{r} f_{ij}$ para cada j = 1,2,..., c.

f_{ij} es la frecuencia observada cuando $X = x_i$, $Y = y_j$.

Entonces H(X), H(X/Y) e I(X/Y) se estiman por:

$$\hat{H}(X) = -\sum_{i=1}^{r} \hat{p}(x_i) \ln \hat{p}(x_i)$$
$$= -\sum_{i=1}^{r} \frac{f_{i\bullet}}{N} \ln\left(\frac{f_{i\bullet}}{N}\right)$$

$$\hat{H}(X/Y) = -\sum_{j=1}^{c}\sum_{i=1}^{r} \hat{p}(x_i, y_j) \ln \hat{p}(x_i / y_j)$$

$$= -\sum_{j=1}^{c}\sum_{i=1}^{r} \frac{f_{ij}}{N} \ln\left(\frac{f_{ij}}{f_{\bullet j}}\right)$$

$$\hat{I}(X/Y) = \sum_i \sum_j \left(\frac{f_{ij}}{N}\right) \log\left(\frac{f_{ij} N}{f_{\bullet j} f_{i\bullet}}\right)$$

$$= \frac{1}{N} \sum_i \sum_j (f_{ij}) \log\left(\frac{f_{ij} N}{f_{\bullet j} f_{i\bullet}}\right)$$

En Kullback (págs. 156, 157) puede encontrarse la demostración de que la hipótesis de independencia entre X y Y puede expresarse como:

Ho: I(X/Y) = 0 (independencia)

Ha: I(X/Y) > 0;

y Kullback demuestra que bajo Ho, la estadística

$$T = 2N\hat{I}(X/Y)$$

tiene una distribución aproximadamente chi-cuadrada con (r-1)(c-1) grados de libertad y por tanto puede usarse para probar la hipótesis de independencia.

La prueba puede generalizarse para probar independencia condicional de la siguiente manera.

En el caso de tres variables aleatorias X, Y, Z, se tiene que la información sobre X cuando $Z = z_l$ es

$$H(X / Z = z_l) = -\sum_{j=1}^{c}\sum_{i=1}^{r} p(x_i / z_l) \ln p(x_i / z_l), \qquad l = 1,2,\ldots,d$$

y la información de X dados $Z = z_l$ y $Y = y_j$ es:

$$H(X / Y = y_j, Z = z_l) = -\sum_{i=1}^{d} p(x_i / y_j, z_l) \ln p(x_i / y_j, z_l)$$

$$= \sum_{l=1}^{d}\sum_{j=1}^{c}\sum_{i=1}^{r} p(x_i, y_j, z_l) \ln p(x_i / y_j, z_l)$$

entonces la información media sobre X dada por Y y Z se define como

$$H(Z / Y, Z) = \sum_{l=1}^{d}\sum_{j=1}^{c}\sum_{i=1}^{r} p(x_i, y_j, z_l) \ln p(x_i / y_j, z_l)$$

la ganancia en información acerca de X por Y dado que se conoce Z es:

$$I(X/Z,Y) = H(X/Z) - H(X/Z,Y).$$

La hipótesis de independencia de X y Y dado Z $(X \perp\!\!\!\perp Y / Z)$ se puede expresar como:

Ho: $I(X/Z,Y) = 0$ (independencia)

Ha: $I(X/Z,Y) > 0$;

y Kullback demuestra que bajo Ho, la estadística

$$T = 2N\hat{I}(X/Z,Y)$$

tiene una distribución chi-cuadrada con (r-1)(c-1)d grados de libertad.

6 APLICACIONES.

6.1 Variables asociadas con la sobrevivencia al infarto agudo al miocardio

Se pretende estudiar el efecto en la sobrevivencia al infarto agudo del miocardio. El estudio incluye N=96 pacientes que ingresaron con infarto agudo al miocardio a la unidad de terapia intensiva del Hospital General de Zona No. 11 del IMSS, durante los años 1991 a junio de 1993. La variable dependiente fue Mortalidad (C = 1 si el paciente sobrevivio a la terapia intensiva, y C = 2 si falleció durante la terapia intensiva).

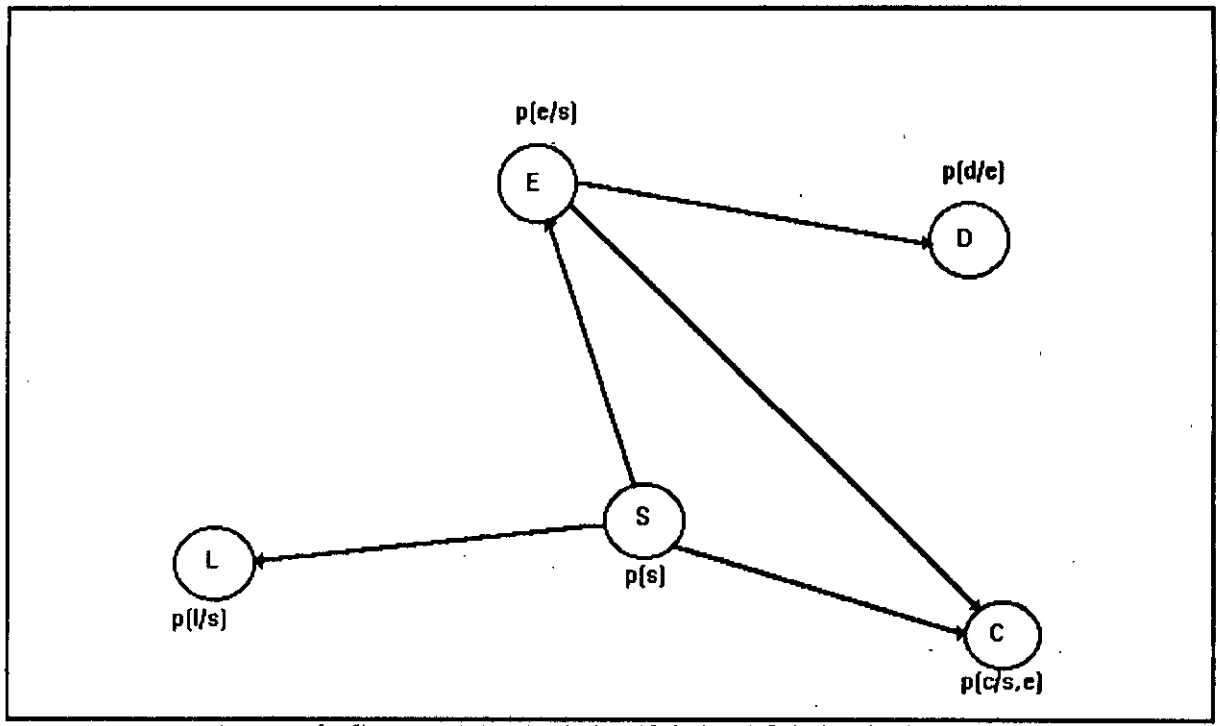
Se midieron 36 variables a cada uno de los pacientes, de las cuales solo 5 serán analizadas.

VARIABLES DE ESTUDIO.

$$\text{EDAD} = \begin{cases} E \leq 55 \\ 55 < E \leq 65 \\ E > 65 \end{cases} \qquad \text{SEXO} = \begin{cases} \text{MASCULINO} \\ \textit{FEMENINO} \end{cases} \qquad \text{DIABETICO} = \begin{cases} \text{SI} \\ \textit{NO} \end{cases}$$

$$\text{MORTALIDAD} = \begin{cases} \text{SI FALLECIO} \\ \textit{NO FALLECIO} \end{cases} \qquad \text{LOCALIZACION } \textit{DEL INFARTO} = \begin{cases} \text{ANTEXT} \\ \text{ANTLAT} \\ \text{ANTSEP} \\ \text{DIAF} \\ \text{POSTINF} \\ \textit{SUB} \end{cases}$$

Los efectos hipotéticos de las variables independientes (E, S, D, L) sobre la variable dependiente C se representa en el siguiente diagrama.



$$p(c,e,s,l,d) \cdot = \cdot p(s)p(e/s)p(l/s)p(d/e)p(c/s,e)$$

DIAGRAMA DE INFLUENCIA

Esta topología nos proporciona información directa sobre las relaciones de dependencia que consideramos entre las variables involucradas.

Tomando como base la topología anterior:

1.- Para el primer par de variables.

$$C = X = \text{Mortalidad} \qquad VS \qquad E = Y = \text{Edad}$$

La hipótesis a probar es:

Ho: La Mortalidad es independiente de la Edad $(C \amalg E)$

VS

Ha: La mortalidad depende de la edad $(C \not\amalg E)$

Primero obtenemos la información incondicional estimada para la variable mortalidad:

$$\hat{H}(x) = 0.51169557.$$

La información media sobre mortalidad proporcionada por la Edad es estimada por:

$$\hat{H}(X / Y) = 0.459043$$

Por lo tanto la ganancia estimada resulta ser:

$$\hat{I}(X / Y) = \hat{H}(X) - \hat{H}(X / Y)$$

$$= 0.05263.$$

y el valor de la estadística es:

$$T = 2N\hat{I} = 2(96)(0.05263)$$

$$= 10.1088$$

y el valor de tablas con α=0.05 con 2 grados de libertad es $X^2_{0.05(2)} = 5.99$

Por lo tanto se rechaza Ho, esto es; existe evidencia suficiente para decir que la Edad está asociada con la Mortalidad , esto se puede observar en la siguiente tabla, la cual muestra que en la categoría (1-Personas jóvenes) de Edad existe menor probabilidad de morir mientras que en las otras dos categorias la probabilidad es mayor.

EDAD	MORTALIDAD	
	NO	SI
1	0.9655	0.0345
2	0.7586	0.2414
3	0.6842	0.3158

TABLA 2. Probabilidades condicionales de Mortalidad dada la Edad.

2.- Para el segundo par de variables

C = X = Mortalidad VS S = Y = Sexo

La hipótesis se plantea como en el caso anterior, solo sustituimos la variable Edad por la variable Sexo. En notación tenemos:

Ho: C⊔S VS Ha: C⊮S

Como la información sobre mortalidad ya la tenemos, entonces solo calcularemos la información sobre mortalidad dada por el sexo.

$$H(X/Y) = 0.486697$$

La ganancia es

$$I(X/Y) = 0.02499857$$

entonces el valor de X^2 calculado es

$$2\hat{I} = 4.7997$$

y el valor de tablas con α=0.05 con 1 grado de libertad resulta $X^2_{0.05(1)} = 3.84$

Como el valor calculado es mayor que el valor de tablas rechazamos Ho, entonces concluimos que la Mortalidad aparentemente esta asociada con el sexo.

Para probar las hipótesis

Ho: C⊔D VS Ha: C⊮D y

Ho: C⊔L VS Ha: C⊮L

se realizaron los mismos pasos que en el caso anterior. Los resultados finales de estas hipótesis se muestran en la Tabla 2.

Cabe mencionar, que las hipótesis que nos muestran dependencia son las que reportaron mayor ganancia de información, por lo tanto, las hipótesis siguientes se prueban condicionando por la variable (o grupo de variables) con I = H(X) - H(X/Y) mayor. En este caso condicionaremos por las variables Edad y Sexo.

Para la primera hipótesis tenemos:

Ho: La Mortalidad es independiente del Sexo dada la Edad. ($C \amalg S / E$)

Ha: La Mortalidad es dependiente del Sexo dada la Edad..($C \not\amalg S / E$)

Primero calculamos H(X/Z,Y) para cada categoria de la variable condicionante. En este caso la condicionante es la Edad, por lo tanto cuando

Edad = 1

H(X/Z,Y) = 0.1433968849

Edad = 2

H(X/Z,Y) = 0.41478936

Edad = 3

H(X/Z,Y) = 0.6213664219

Ahora realizamos la sumatoria de multiplicar la probabilidad marginal de cada categoria por H(X/Z,Y).

$$\sum_{i=1}^{3} p(z_i) H(X / Z_i, Y) = 0.41456489$$

Y como ya conocemos la información de mortalidad dada la Edad se tiene

H(X/Z) = 0.459043

entonces la información resulta

$$\hat{I} = 0.459043 - 0.41456489 = 0.44478; \quad \text{por lo tanto}$$

$$T = 4.2699.$$

Este valor lo comparamos con un valor de tablas X^2 con $\alpha = 0.05$, con (r-1)(c-1)d grados de libertad; donde d son las categorias de la variable condicionante, entonces el valor de tablas es igual a 7.81.

Como el valor de tablas es mayor que el valor calculado, entonces no rechazamos Ho, por lo tanto existe independencia entre Mortalidad y Sexo dada la Edad.

Del mismo modo se prueban las hipótesis restantes, estas se muestran junto con los resultados finales en la Tabla 2.

HIPOTESIS Ho	$\hat{I}$	Valor Crítico X^2 (α= 0.05)	DECISION
C ⫫ E	10.1088	5.99*	SE RECHAZA Ho
C ⫫ S	4.7997	3.84*	SE RECHAZA Ho
C ⫫ D	0.8789	3.84	NO SE RECHAZA Ho
C ⫫ L	5.1605	11.07	NO SE RECHAZA Ho
C ⫫ S / E	4.2699	7.81	NO SE RECHAZA Ho
C ⫫ D / E	0.9752	7.81	NO SE RECHAZA Ho
C ⫫ L / E	6.7753	25.00	NO SE RECHAZA Ho
C ⫫ D / S	0.3696	5.99	NO SE RECHAZA Ho
C ⫫ L / S	3.6818	18.3	NO SE RECHAZA Ho

TABLA 3. Resultados del ejemplo Infarto.

De esta tabla concluimos que la Mortalidad de los pacientes diabéticos depende de la Edad que tengan estos; así como del Sexo, pero esta última variable no tiene un efecto significativo dada la variable Edad.

Esto se puede apreciar de forma resumida y clara en el Diagrama de Influencia Probabilista que se muestra a continuación.

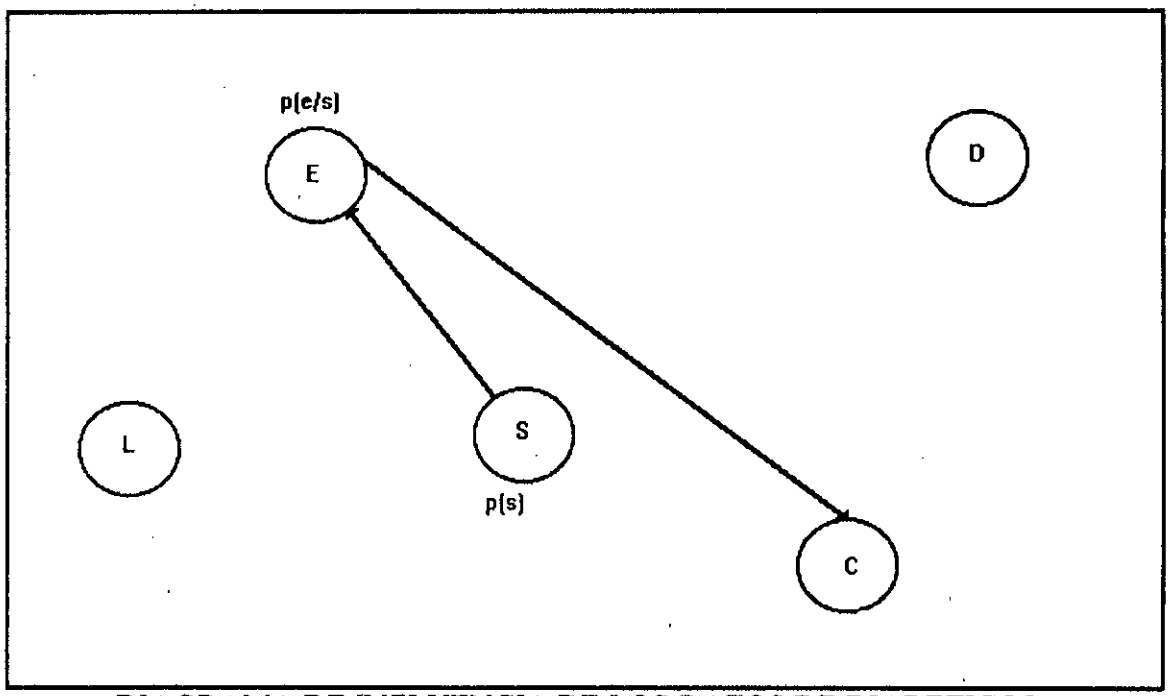


DIAGRAMA DE INFLUENCIA DE LOS DATOS DE DIABETICOS.

6.2 Variables socieconómicas relacionadas con las preferencias electorales.

A continuación presentamos otra ejemplificación del uso de estas técnicas.

Se quiere conocer la opinión de la ciudadanía Veracruzana sobre sus preferencias electorales. En este caso utilizamos 5 variables de estudio.

Dedica
- 1 Campo.
- 2 Otros Servicios.
- 3 Independiente.
- 4 Desempleado.

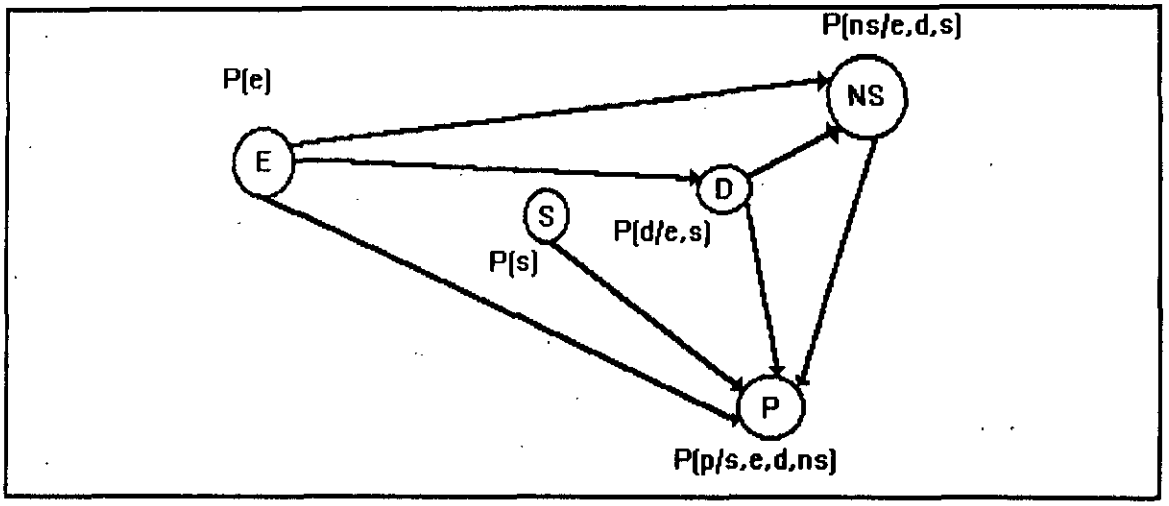
Presi
- 1 PRI
- 2 PAN
- 3 PRD
- 4 No saben si van a votar.
- 5 Otros.

Edad
- 1 (18-30)
- 2 (31-45)
- 3 (45- má s)

Sexo
- 1 Hombres
- 2 Mujeres

Nivel Socioeconómico
- 1 Baja
- 2 Media
- 3 Alta

Para las variables la topología que consideramos es la siguiente:



p(e,s,d,ns,p)=p(e)p(s)p(d/e,s)p(ns/e,d,s)p(p/s,e,d,ns)

DIAGRAMA DE INFLUENCIA

Tomando como base las conecciones de los nodos que se presentan en la gráfica, se realizan los cálculos necesarios para probar las hipótesis de dos variables, las cuales se presentan en la Tabla 4 con sus respectivos resultados.

Es importante señalar que las hipótesis que nos muestran dependencia, las que reportan mayor ganancia de información son

$$\hat{I}(P / NS) = 0.011089$$

$$\hat{I}(P / D) = 0.00871417$$

por lo tanto, las hipótesis se probaran condicionando por estas dos variables y los resultados se muestran en la Tabla 4.

HIPOTESIS Ho	$\hat{I}$.	Valor Crítico X^2 (α= 0.05)	DECISION
D ⫫ NS	105.9785	12.59*	SE RECHAZA Ho
D ⫫ E	72.688	12.59*	SE RECHAZA Ho
D ⫫ S	175.487	7.81*	SE RECHAZA Ho
P ⫫ D	29.3145	21.03*	SE RECHAZA Ho
P ⫫ S	11.7450	9.49*	SE RECHAZA Ho
P ⫫ NS	37.3050	15.51*	SE RECHAZA Ho
P ⫫ E	21.1606	15.51*	SE RECHAZA Ho
P ⫫ S / NS	9.3180	21.03	NO SE RECHAZA Ho
P ⫫ D / NS	25.0040	43.77	NO SE RECHAZA Ho
P ⫫ E / NS	39.5737	36.415	NO SE RECHAZA Ho
P ⫫ S / D	16.4110	26.296	NO SE RECHAZA Ho
P ⫫ NS / D	30.0700	46.194	NO SE RECHAZA Ho
P ⫫ E / D	23.1910	46.194	NO SE RECHAZA Ho

TABLA 4. Resultados del ejemplo de Opinión de la Ciudadanía Veracruzana.

De ésta tabla se deduce, que el porcentaje de votantes por determinado partido depende basicamente de el nivel socieconómico

Graficamente se tiene

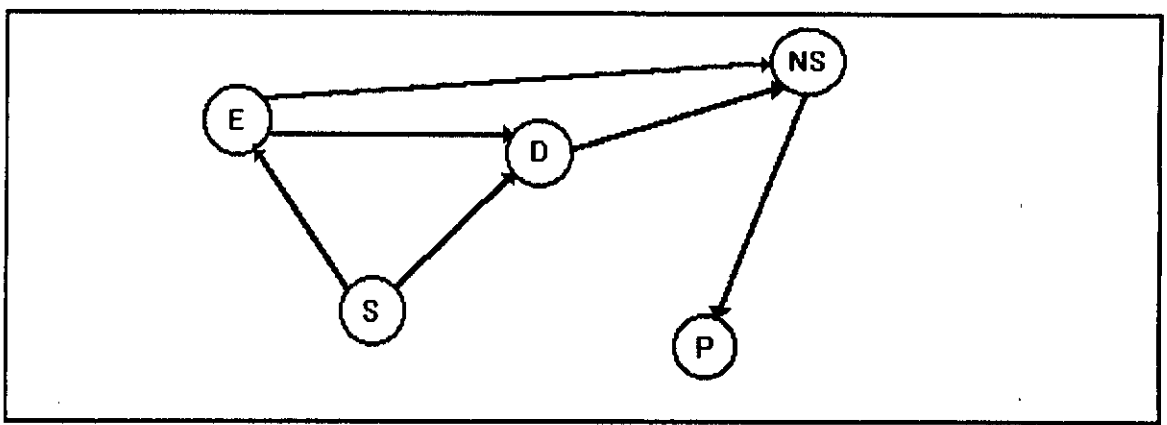


DIAGRAMA DE INFLUENCIA DE LOS DATOS DEL ESTUDIO DE OPINION DE LA CIUDADANIA VERACRUZANA.

En la siguiente tabla se presentan las probabilidades condicionales de votar por un partido dado que se conoce el nivel socieconómico P(Partido/NS), y las probabilidades condicionales agregando la variable D (ocupación). Como puede observarse, una vez que se conoce NS, el conocimiento de D no modifica sustancialmente las probabilidades condicionales, como es de esperarce por las pruebas de independencia condicional. (Donde se notan más fluctuaciones es debido a que el número de casos es muy pequeño).

		P(P / NS, D)			
		D = 1	D = 2	D = 3	D = 4
P(PRI/NS=1)	0.46	0.48	0.47	0.41	0.27
P(PRD/NS=1)	0.18	0.23	0.16	0.17	0.37
P(PAN/NS=1)	0.14	0.12	0.14	0.20	0.09
P(PRI/NS=2)	0.38	0.35	0.38	0.42	0.35
P(PRD/NS=2)	0.14	0.23	0.14	0.15	0.18
P(PAN/NS=2)	0.26	0.13	0.27	0.21	0.12
P(PRI/NS=3)	0.44	0.66	0.43	0.41	0.60
P(PRD/NS=3)	0.13	0.33	0.11	0.17	0.00
P(PAN/NS=3)	0.24	0.00	0.23	0.32	0.120

TABLA 5. Probabilidades condicionales de preferencia electoral sobre el nivel socioeconómico y ocupación.

REFERENCIAS

Agresti, A. (1984). Categorical Data Analysis. Wiley N.Y.

Barlow, R.E. and Braganca-Pereira, C.A. (1990). Conditional Independence and Probabilistic Influence Diagrams. En Topics in Statistical Dependence. Lecture Notes-Monograph Series. Institute of Mathematical Statistics.

Dawid, A:P: (1979). Conditional Independence in Statistical Theory. J.R. Statist. Soc. B. 41, 1-31.

Kullback, S. (1968). Information Theory and Statistics. Dover Pub. Inc. New York.

L. Sucar. (1993). Structure and Parameter Learning in Probabilistic Networks. Centro de Ing. Computacional ITESM-Campus Morelos. Cuernavaca, Morelos.

Martínez, M., Montano, A. y Ojeda, M. M. (1992). Independencia Condicional y Diagramas de Influencia Probabilista; Dos Aplicaciones. Memorias del VII FORO NACIONAL DE ESTADISTICA CHolula, Puebla.

Montano, A. y Ojeda, M. M. (1991). Análisis Estadístico del Efecto de Sustratos y Fertilización en Crisantemos. Reporte Interno, LINAE; Fac. de Estadística U.V. Xalapa,Ver.

Ojeda, M. M. (1992). Modelos Lineales Generalizados para Analizar Datos en Grupos. Revista Investigación Operacional. Vol. 13 No. 2. pp.140 , 147.