2; 1

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

MEDIDAS DE ASOCIACION PARA VARIABLES NOMINALES

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I A

P R E S E N T A

REBECA AGUIRRE HERNANDEZ

MEXICO, D. F. 1988

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

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

INTRODUCCION

El objetivo de esta tesis es presentar una recopilación de distintas medidas de asociación para variables nominales que han sido publicadas en diversos libros y revistas, y presentarlas de tal forma que una persona que tenga conocimientos de estadística elemental pueda saber qué medidas usar en un problema particular. Este trabajo se llevó a cabo porque una parte importante en el estudio de tablas de contingencia es el análisis, por medio de índices, del grado de asociación entre las variables involucradas en la tabla.

Cada Capítulo reune diversas medidas de asociación que tienen características similares o que fueron propuestas con fines semejantes. Casi todos los índices o medidas de asociación se presentan de acuerdo con el patrón que se indica a continuación:

1. Definición del índice poblacional.
2. Descripción y demostración de algunas de sus propiedades.
3. Interpretación del índice poblacional y, en algunos casos, ilustración de dicha interpretación mediante un ejemplo.
4. Mención de sus ventajas o desventajas.
5. Enunciación de relaciones entre las medidas de asociación, y

6. Expresión del estimador de máxima verasimilitud del índice poblacional y la varianza o desviación estándar asintótica de dicho estimador.

Esta tesis está integrada por seis Capítulos y un Apéndice; al final de cada uno de ellos se menciona la bibliografía consultada.

El Capítulo I tiene como objetivo definir y ejemplificar algunos conceptos básicos en estadística que se manejarán en los siguientes capítulos. Entre otros términos se definen las medidas de asociación y la asociación perfecta entre dos variables nominales.

El Capítulo II trata acerca del coeficiente de correlación y las medidas de asociación basadas en la estadística ji-cuadrada.

En el Capítulo III se describen las medidas de asociación basadas en el cociente de productos cruzados.

En el Capítulo IV se discuten varias medidas de asociación basadas en la lógica de reducción proporcional en el error. Algunos de los índices propuestos para dos variables nominales se generalizan para medir la asociación entre tres variables nominales.

El Capítulo V, a diferencia de los demás, reune varias medidas de asociación distintas entre sí. Algunas de estas medidas son muy antiguas, otras están basadas en el concepto de estructuras latentes o en el de disimilaridad y algunas son -

para problemas muy particulares.

En el Capítulo VI se discute un grupo de índices que miden un tipo especial de asociación: la confiabilidad.

Por último en el Apéndice se explica cómo funcionan los programas de cómputo: "RUENTAJA" y "DATCUALIT" que calculan algunas medidas de asociación para variables nominales.

I. CONCEPTOS BASICOS

1. Variables aleatorias.
2. Escalas de medición.
3. La población objetivo y la muestra.
4. Tablas de contingencia.
5. Esquemas de muestreo para tablas de contingencia.
6. Estimación de los parámetros en una tabla de contingencia.
7. Independencia en una tabla de contingencia.
8. Conceptos generales acerca de medidas de asociación.
9. El coeficiente de correlación de Pearson.
10. Bibliografía.

I. CONCEPTOS GENERALES

Un método que emplea la ciencia para predecir y describir un fenómeno natural es la construcción de modelos matemáticos - que representen de manera adecuada al fenómeno en estudio. Un - tipo especial de modelos son los probabilísticos. Debido a que estos modelos se usarán a lo largo de esta tesis es necesario - mencionar algunos conceptos relacionados con ellos.

1. VARIABLES ALEATORIAS.

Los fenómenos naturales pueden ser conceptualizados en la teoría de la probabilidad como un experimento que puede conducir a distintos resultados. Por ejemplo, el lanzar una moneda se puede concebir como un experimento con únicamente dos posibles resultados: que caiga en águila o que caiga en sol. Al - conjunto de todos los posibles resultados de un experimento se le denomina espacio muestral. Los resultados de un experimento pueden ser numéricos (el número de microorganismos vivos en un cultivo, las ventas semanales de una empresa, el peso de un - animal, etc.) o no numéricos (el color de ojos, la cara de una moneda, la clase social, etc.) y para facilitar su análisis a cada resultado se le asigna un número; este es el concepto de - variable aleatoria.

1.1 Definición.

Una variable aleatoria es una función que asigna un número real a cada elemento del espacio muestral. Las variables aleatorias generalmente se denotan con letras mayúsculas: A, B, C,

W, X, Y, Z, ... En el ejemplo de la moneda, se puede definir una variable aleatoria X que represente el número de soles que se obtienen en un lanzamiento. En este caso, si la moneda cae en águila X=0 y si cae en sol X=1. Como se puede ver, a los elementos del espacio muestral, águila y sol, la variable aleatoria X les asignó, respectivamente, el valor de cero y uno.

1.2 Variables Dicotómicas.

Las variables dicotómicas únicamente toman dos valores distintos. Un ejemplo es la variable aleatoria X definida hace un momento.

1.3 Variables Cualitativas y Cuantitativas.

Las variables aleatorias se dividen en cualitativas y cuantitativas. Las primeras codifican o clasifican los resultados de un experimento y por lo general asignan números a los espacios muestrales integrados por elementos no numéricos. Cuando tiene sentido realizar operaciones aritméticas entre los valores de una variable aleatoria se dice que dicha variable es cuantitativa. Generalmente estas variables asignan números a espacios muestrales formados por elementos numéricos.

Dependiendo de la escala usada para registrar los resultados de un experimento, las variables cualitativas se dividen en nominales y ordinales y las variables cuantitativas se clasifican como de intervalos o de razón. La Sección 2 trata acerca de las escalas de medición.

1.4 Variables Simétricas y Asimétricas.

En ocasiones al estudiar dos o más variables a la vez se puede establecer una relación de causa-efecto entre ellas, se pueden jerarquizar de acuerdo al orden en el que fueron observadas o se pueden clasificar como variables controladas y no controladas. En estos casos la relación entre las variables es asimétrica. Las variables que producen un cierto efecto en el resto, que preceden cronológicamente a las otras o que pueden ser controladas se conocen como variables explicativas o independientes; las demás reciben el nombre de variables respuesta o dependientes. Por otra parte, la relación entre dos o más variables es simétrica cuando ninguna de ellas precede a las otras en algún sentido.

2. ESCALAS DE MEDICION.

Con el fin de comprender un fenómeno muchas veces es necesario evaluar o medir algunas de sus propiedades más relevantes. Existen varias escalas de medición pero las más conocidas son: nominal, ordinal, de intervalos y de razón. Para cada escala hay distintos procedimientos estadísticos que permiten realizar un análisis de los datos registrados.

2.1 Escala Nominal.

En cualquier ciencia, la operación más simple es la de clasificar. Clasificar es agrupar, respecto a una determinada característica, a un conjunto de individuos estableciendo las diferencias y las semejanzas entre ellos con el fin de que las

categorías o grupos que se obtengan sean lo más homogéneos posible. La clasificación es el nivel más simple de medición y se le conoce como escala nominal. Al efectuar una clasificación, las categorías obtenidas son etiquetadas con algún nombre. Por ejemplo, al clasificar a un grupo de personas de acuerdo a su religión se pueden definir las categorías: hinduista, budista, judaista, islamista y cristiano o bien, las categorías 0, 1, 2, 3 y 4. Si un individuo pertenece a la categoría de los cristianos o a la categoría 3 no implica que, en algún sentido, sea "mejor que" o "mayor que" otro individuo clasificado en una categoría distinta. Es importante señalar que cuando las categorías son etiquetadas con números no tiene sentido realizar operaciones aritméticas entre ellos.

En una escala nominal la relación entre dos o más individuos puede ser simétrica y transitiva. Un ejemplo de este tipo de relación es la igualdad (=).

Para poder realizar un análisis estadístico con base en una clasificación es necesario que las categorías sean exhaustivas (es decir, se deben considerar todos los posibles casos de clasificación) y además deben ser mutuamente excluyentes (esto significa que ningún individuo debe quedar clasificado en dos o más categorías a la vez).

2.2 Escala Ordinal.

En ocasiones, las categorías reflejan en forma aproximada con qué intensidad poseen los individuos la característica en -

estudio; en estos casos existe un orden implícito entre las categorías y la escala de medición se conoce como escala ordinal. Un ejemplo es la clasificación de un conjunto de familias de acuerdo a su nivel socioeconómico como: clase baja, clase media-baja, clase media-alta y clase alta. Las familias clasificadas en la clase media-alta tienen un mejor nivel socioeconómico que las clasificadas en la clase baja; sin embargo, no se sabe qué tan grande es la diferencia entre el nivel socioeconómico de unas y otras familias.

Al trabajar con una escala ordinal se pueden establecer relaciones asimétricas y transitivas entre los individuos en estudio. Un ejemplo de este tipo de relación es la de "mayor que (>)".

2.3 Escala de Intervalos.

La escala de intervalos está siempre asociada con alguna unidad física de medición. Por esto, al trabajar con dicha escala, tiene sentido sumar o restar los datos. Un ejemplo típico es la medición de la temperatura, ya sea en grados Farenheit o centígrados. En este caso se pueden hacer afirmaciones como la siguiente: la diferencia entre los 70°F y los 35°F es la misma que existe entre los 105°F y los 70°F. Sin embargo, en la escala de grados Farenheit o centígrados el cero es un punto arbitrario y por esto es incorrecto decir que un objeto cuya temperatura es de 40°C es dos veces más caliente que uno con una temperatura de 20°C.

2.4 Escala de Razón.

Una escala de razón es aquella en la cual se puede localizar un punto no arbitrario que represente la ausencia de la característica en estudio. En esta escala, los datos pueden ser comparados por medio de su cociente. Algunas características que se pueden medir mediante una escala de razón son la longitud (en metros), el peso (en gramos o libras), el ingreso (en pesos), el tiempo (en segundos), etc.

3. POBLACION Y MUESTRA.

Uno de los objetivos de la estadística es hacer inducciones o generalizaciones acerca de una población con base en un subconjunto de individuos de dicha población. Un ejemplo es el siguiente: se tienen almacenadas 10 millones de semillas que producen flores blancas o rojas y antes de venderlas se quiere saber cuántas (o que porcentaje) de las semillas producirán flores blancas. Con este fin se siembran unas cuantas semillas y de acuerdo al color de sus flores se predice el color de las flores del resto de las semillas. Al proceder de esta manera no se puede saber exactamente cuántas semillas producirán flores blancas, pero si las semillas que se siembran son seleccionadas mediante un procedimiento especial, se podrá hacer una afirmación probabilística para las semillas restantes; es decir los resultados obtenidos con base en unas cuantas semillas se pueden generalizar con una cierta confianza de que la generalización sea correcta.

Este ejemplo ilustra cómo se pueden adquirir conocimientos sobre una población determinada cuando se tiene información parcial de ella. Las razones por las cuales no siempre se puede estudiar a toda la población son:

1. Estudiar a cada uno de sus elementos puede ser costoso y tardado.
2. La población puede ser infinita o durante su estudio puede presentar algún cambio (esto es lo que sucede con las semillas).

En seguida se define qué es la población objetivo y qué es una muestra.

3.1 Definición: Población Objetivo.

La totalidad de los individuos sobre los que se quiere adquirir información se conoce como población objetivo. La población objetivo se denomina también población de interés o población en estudio.

3.2 Definición: Muestra.

Una muestra es un subconjunto de individuos de la población objetivo.

En el ejemplo anterior, los 10 millones de semillas almacenadas constituyen la población objetivo y las semillas que se siembran son la muestra. En estadística solo se pueden hacer inferencias cuando se conoce la probabilidad de que un individuo de la población objetivo sea incluido en la muestra. A

estas muestras se les denominará, en este trabajo, como muestras aleatorias.

4. TABLAS DE CONTINGENCIA.

En las ciencias sociales y biológicas es común clasificar a un conjunto de individuos respecto a dos o más variables cualitativas. Los datos así obtenidos se pueden presentar en una tabla de conteos que recibe el nombre de tabla de contingencia. Por ejemplo, en la tabla siguiente (Steel y Torrie, 1984) se puede observar cómo fueron clasificados 77 pinos blancos de acuerdo a su clase de edad y a la reacción que presentaron al ser infectados con hongos.

Tabla 4.1

		Edad del Pino (años)				
		[4, 10)	[10, 20)	[20, 40)	≥ 40	Totales
Reacción	Sano	7	6	11	15	39
	Enfermo	14	11	5	8	38
	Totales	21	17	16	23	77

Esta tabla solo exhibe la clasificación de los 77 pinos observados y no la clasificación de toda la población de interés. Para la población objetivo se puede diseñar una tabla de contingencia cuyos elementos sean constantes desconocidas que representen la probabilidad de que un individuo de la población,

seleccionado al azar, tenga un determinado conjunto de características.

4.1 Tabla de Contingencia para la Población.

La tabla de contingencia para la población objetivo se representará como:

Tabla 4.1.1

		B						
		B_1		B_j		B_J	Tot.
A	A_1	p_{11}		p_{1j}		p_{1J}	$p_{1.}$
	A_i	p_{i1}		p_{ij}		p_{iJ}	$p_{i.}$
	A_I	p_{I1}		p_{Ij}		p_{IJ}	$p_{I.}$
	Tot.	$p_{.1}$		$p_{.j}$		$p_{.J}$	1

En esta tabla, A y B denotan las variables en estudio. La variable A tiene I categorías o clases: $A_1, \ldots, A_i, \ldots, A_I$ y la variable B tiene J categorías: $B_1, \ldots, B_j, \ldots, B_J$.; por lo tanto, la tabla anterior es de dimensión $I \times J$. Un individuo clasificado en A_i y B_j pertenece al i-ésimo renglón y a la j-ésima columna de la tabla 4.1.1; es decir, pertenece a la celda (A_i, B_j). La probabilidad de que un individuo esté clasificado en dicha celda se denota como p_{ij}.

Además, la probabilidad de que un individuo de la población,

seleccionado al azar, pertenezca al i-ésimo renglón de la ta-- bla 4.1.1 es:

$$p_{i1} + \ldots + p_{ij} + \ldots + p_{iJ} = \sum_j p_{ij} = p_{i.}$$

ésta es la probabilidad marginal de A_i. Análogamente, la probabilidad de que un individuo esté clasificado en B_j es:

$$p_{1j} + \cdots + p_{ij} + \cdots + p_{Ij} = \sum_i p_{ij} = p_{.j}$$

ésta es la probabilidad marginal de B_j. La suma de todas las - probabilidades marginales de A y B debe ser igual a uno. Sim-- bólicamente:

$$p_{1.} + \ldots + p_{i.} + \ldots + p_{I.} = \sum_i p_{i.} = 1$$

$$\text{y} \quad p_{.1} + \ldots + p_{.j} + \ldots + p_{.J} = \sum_j p_{.j} = 1$$

En la mayoría de los problemas las p_{ij} son desconocidas; su valor se estima a partir de una muestra.

4.2 Tabla de Contingencia para la Muestra.

La tabla de contingencia para la muestra es:

Tabla 4.2.1

		B				
		B_1		B_j		B_J	Tot.
A	A_1	N_{11}		N_{1j}		N_{1J}	$N_{1.}$
	A_i	N_{i1}		N_{ij}		N_{iJ}	$N_{i.}$
	A_I	N_{I1}		N_{Ij}		N_{IJ}	$N_{I.}$
	Tot.	$N_{.1}$		$N_{.j}$		$N_{.J}$	n

Aquí, N_{ij} es una variable aleatoria que denota el número

de individuos de la muestra que serán clasificados en las categorías A_i y B_j. El número de individuos que serán clasificados en A_i es $N_{i.}$ A $N_{i.}$ se le conoce como la frecuencia marginal de A_i y es igual a $N_{i1} + \ldots + N_{ij} + \ldots + N_{iJ} = \sum_j N_{ij}$.

La frecuencia marginal de B_j es:

$$N_{.j} = \sum_i N_{ij} = N_{1j} + \ldots + N_{ij} + \ldots + N_{Ij} \quad \forall\ j = 1, \ldots, J .$$

La suma de todas las frecuencias marginales de A, así como las de B, debe ser igual a n, el tamaño de la muestra.

Regresando a la tabla 4.1., se puede ver que en ella se clasificó a una muestra de 77 pinos blancos de acuerdo a los variables:

A: reacción que presentaron los pinos al ser infectados con hongos

y B: edad de los pinos.

Dicha tabla es de dimensión 2 X 4 ya que I = 2 y J = 4.

Las categorías de A y B son:

A_1 : Sano $\quad B_1$: $[4, 10)$

A_2 : Enfermo $\quad B_2$: $[10, 20)$

B_3 : $[20, 40)$

B_4 : ≥ 40

Las frecuencias marginales de A_1 y A_2, 39 y 38, indican que al finalizar el experimento hubo aproximadamente el mismo número de pinos sanos y enfermos. Las frecuencias marginales de B_1, B_2, B_3 y B_4, $N_{.1}$, $N_{.2}$, $N_{.3}$ y $N_{.4}$ con iguales a 21, 17, 16 y 23 respectivamente. Además, la celda (A_1, B_4) es

la que mayor número de observaciones tiene y la celda (A_2, B_3) es la que cuenta con el menor número de datos.

5. ESQUEMAS DE MUESTREO.

En la Sección 4.2. se dijo que, en la tabla de contingencia muestral, Nij es una variable aleatoria, pero no se mencionó nada acerca del tamaño de la muestra o de las frecuencias marginales. A continuación se verá que n y las frecuencias marginales de A y B pueden o no ser aleatorias, todo depende del problema en estudio.

5.1 El tamaño de la Muestra es Fijo.

En ocasiones el número de individuos que va a ser observado se determina antes de llevar a cabo el estudio. Cuando esto sucede se dice que en la tabla 4.2.1, el tamaño de la muestra es fijo y las frecuencias marginales de A y B son variables aleatorias. Además, en este caso las variables Nij se distribuyen como una multinomial ya que la probabilidad de clasificar Oij individuos en la celda (A_i, B_j), $\forall$ i,j es igual a:

$$f_{N_{11}\ldots N_{IJ}}(O_{11},\ldots,O_{IJ};p_{11},\ldots,p_{IJ}) = \frac{n!}{\prod_i \prod_j O_{ij}!} \prod_i \prod_j p_{ij}^{O_{ij}}$$

Esta es la función de densidad multinomial; n es el tamaño de la muestra, Oij es un número entero que denota un posible valor de Nij y pij es la probabilidad de que un individuo de la población pertenezca a la celda (A_i, B_j).

5.2 Las Frecuencias Marginales de una de las Variables están Fijas.

En algunos estudios el objetivo es comparar el comportamiento de dos o más poblaciones respecto a una variable cualitativa A. En este tipo de problemas los datos recabados al observar una muestra de cada población se pueden presentar en una tabla de contingencia en la cual las categorías de B se refieran a las poblaciones de interés; por ejemplo:

B: Poblaciones

		Pob. 1		Pob. j		Pob. J	Tot.
	A_1	N_{11}		N_{1j}		N_{1J}	$N_{1.}$
A	A_i	N_{i1}		N_{ij}		N_{iJ}	$N_{i.}$
	A_I	N_{I1}		N_{Ij}		N_{IJ}	$N_{I.}$
	Tot.	$m_{.1}$		$m_{.j}$		$m_{.J}$	n

La variable Nij denota el número de individuos de la población j que serán clasificados en la i-ésima categoría de A y las frecuencias marginales de B son el tamaño de la muestra extraida de cada población; es decir, m.j es el número de individuos muestreados de la población j. Si antes de realizar el estudio se establece el tamaño de la muestra para cada población entonces las frecuencias marginales de B no son

aleatorias pero si las de A.

Cuando se tienen J poblaciones y a partir de cada una de ellas se extrae una muestra de manera independiente usando un esquema de muestreo multinomial entonces la distribución conjunta de $N_{11}, \ldots, N_{ij}, \ldots N_{IJ}$ es una multinomial producto.

5.3 Otros Esquemas de Muestreo.

En biología a veces se trabaja con tablas de contingencia en las que el tamaño de la muestra así como las frecuencias marginales de A y B son variables aleatorias. Este método de muestreo se conoce como Poisson.

También hay tablas en las que todas las frecuencias marginales son fijas (es decir, ninguna de ellas es aleatoria). Este tipo de tablas son poco comunes. Un ejemplo se presenta en el libro de Kendall (1963).

Los resultados sobre estimación que se mencionan en los próximos Capítulos están basados en el supuesto de que el esquema de muestreo es multinomial o multinomial producto.

6. ESTIMACION.

Más adelante se verá que las medidas de asociación son funciones que dependen de las probabilidades poblacionales. Como éstas son desconocidas entonces va a ser necesario estimar las medidas de asociación. Por esta razón, el objetivo de esta sección es enunciar un método de estimación puntual -el de

máxima verosimilitud- y el Principio de Invarianza. El método de máxima verosimilitud servirá para estimar las probabilidades poblacionales y después, aplicando el Principio de Invarianza - se estimarán las medidas de asociación.

6.1 Método de Máxima Verosimilitud.

Los estimadores de máxima verosimilitud se obtienen a partir de la función de verosimilitud que se define a continuación.

6.1.1 Definición: Función de Verosimilitud.

La función de verosimilitud de las variables aleatorias - $X_1, X_2, \ldots, X_n$ es: $f_{X_1 \ldots X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_K)$ la función de densidad conjunta de $X_1, X_2, \ldots, X_n$.

A la función de verosimilitud se le denota como $L(\theta_1,\ldots,\theta_K;x_1,\ldots,x_n)$ para hacer énfasis en que es una función de los K parámetros: $\theta_1,\ldots,\theta_K$. (K es un entero mayor o igual a uno).

6.1.2. Definición: Estimador de Máxima Verosimilitud.

Sea $L(\theta_1,\ldots,\theta_K)=L(\theta_1,\ldots,\theta_K;x_1,\ldots,x_n)$ la función de verosimilitud de las variables aleatorias $X_1, X_2, \ldots, X_n$. Si para - toda $i = 1,\ldots, k$, θ_i pertenece al espacio parametral $\overline{\Theta}$ y - $\hat{Q}_1(x_1,\ldots,x_n),\ldots,\hat{Q}_K(x_1,\ldots,x_n)$ son, respectivamente, los valores de $\theta_1,\ldots,\theta_K$ que maximizan $L(\theta_1,\ldots,\theta_K)$ entonces $\hat{Q}_1(X_1,\ldots,X_n),\ldots,\hat{Q}_K(X_1,\ldots,X_n)$ son los estimadores de máxima verosimilitud de $\theta_1,\ldots,\theta_K$.

En el teorema siguiente se indica cuáles son los estimadores de máxima verosimilitud de las probabilidades poblacionales, pij, suponiendo que el esquema de muestreo es multinomial.

6.1.3 Teorema.

Si la función de densidad conjunta de las variables aleatorias $N_{11}, \ldots, N_{IJ}$ es:

$$f_{N_{11}\ldots N_{IJ}}(O_{11},\ldots,O_{IJ};p_{11},\ldots,p_{IJ})=\frac{n!}{\prod_i\prod_j O_{ij}!}\prod_i\prod_j p_{ij}^{O_{ij}} \quad \text{con } \sum_i\sum_j p_{ij}$$

entonces, los estimadores de máxima verosimilitud de $p_{11}, \ldots,$ pij,..., p_{IJ} son; respectivamente, $\hat{p}_{11}=\frac{N_{11}}{n}, \ldots, \hat{p}_{ij}=\frac{N_{ij}}{n}, \ldots, \hat{p}_{IJ}=\frac{N_{IJ}}{n}$

Demostración.

Por hipótesis, la función de verosimilitud de las variables aleatorias $N_{11}, \ldots, N_{IJ}$ es:

$$L(p_{11},\ldots,p_{IJ})=\frac{n!}{\prod_i\prod_j O_{ij}!}\prod_i\prod_j p_{ij}^{O_{ij}}$$

en donde: $\sum_i\sum_j p_{ij}=1$

Nota:

Para determinar en qué punto se maximiza la función de verosimilitud se calculará el máximo de:

$$\log L(p_{11},\ldots,p_{IJ})=\log\left(\frac{n!}{\prod_i\prod_j O_{ij}!}\right)+\sum_i\sum_j O_{ij}\log p_{ij}$$

ya que es más sencillo trabajar con esta función y ademas su máximo coincide con el de $L(p_{11}, \ldots, p_{IJ})$.

El método que se usará para maximizar la función $\log L(p_{11},\ldots,p_{IJ})$ se conoce como el método de los Multiplicadores de Lagrange.

Las derivadas de $\mathcal{L}(\lambda, p_{11}, \ldots, p_{IJ}) = \log L(p_{11}, \ldots, p_{IJ}) - \lambda\left(\sum_i \sum_j p_{ij} - 1\right)$ respecto a $\lambda, p_{11}, \ldots, p_{IJ}$ son:

$$\frac{\partial \mathcal{L}}{\partial \lambda}(\lambda, \ldots, p_{IJ}) = \sum_i \sum_j p_{ij} - 1 \quad , \quad \frac{\partial \mathcal{L}}{\partial p_{11}}(\lambda, p_{11}, \ldots, p_{IJ}) = \frac{o_{11}}{p_{11}} - \lambda$$

$$\ldots \quad \frac{\partial \mathcal{L}}{\partial p_{IJ}}(\lambda, p_{11}, \ldots, p_{IJ}) = \frac{o_{IJ}}{p_{IJ}} - \lambda$$

Por lo tanto: $\frac{\partial \mathcal{L}}{\partial \lambda}(\lambda, p_{11}, \ldots, p_{IJ}) = 0$ si $\sum_i \sum_j \hat{p}_{ij} = 1$

y $\frac{\partial \mathcal{L}}{\partial p_{ij}}(\lambda, p_{11}, \ldots, p_{IJ}) = 0$ si $\frac{o_{ij}}{\hat{p}_{ij}} = \hat{\lambda} \quad \forall i, j$ (E - 6.1.3.2.)

Al despejar o_{ij} de la ecuación (E — 6.1.3.2) se obtiene que:

$$o_{ij} = \hat{\lambda}\hat{p}_{ij} \quad \Rightarrow \quad \sum_i \sum_j o_{ij} = \hat{\lambda} \sum_i \sum_j \hat{p}_{ij}$$

Como $\sum_i \sum_j o_{ij} = n$ y $\sum_i \sum_j \hat{p}_{ij} = 1$

entonces: $\hat{\lambda} = n$

Si se sustituye $\hat{\lambda}$ en la ecuación (E - 6.1.3.2) y se despeja $\hat{p}_{ij}$ se obtiene que $\hat{p}_{ij} = \frac{o_{ij}}{n}$. Por lo tanto, el estimador de máxima verosimilitud de p_{ij} es:

$$\hat{p}_{ij} = \frac{N_{ij}}{n} \quad \forall \; i = 1, \ldots, I \text{ y } j = 1, \ldots, J \qquad \square$$

Este teorema indica que, si el esquema de muestreo es multinomial, el estimador de máxima verosimilitud de

p_{ij} es la proporción, respecto a n , de individuos de la muestra que serán clasificados en la celda (A_i, B_j).

En el artículo de Birch (1963) se demuestra que, independientemente de que el esquema de muestreo sea multinomial, multinomial producto o Poisson, el estimador de máxima verosimilitud de p_{ij} es $\hat{p}_{ij} = \frac{N_{ij}}{n}$.

A continuación se enuncia la propiedad de invarianza de los estimadores de máxima verosimilitud. Esta propiedad establece bajo que condiciones se puede estimar una función de $p_{11}, \ldots, p_{ij}, \ldots, p_{IJ}$.

6.1.4. Teorema: Principio de Invarianza de los Estimadores de Máxima Verosimilitud (Mood, 1974).

Sea f una función de densidad con parámetro $\underline{\theta} = (\theta_1, \ldots, \theta_K) \in \underline{\Theta}$ y sea $\hat{\underline{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$ con $\hat{\theta}_j = \hat{Q}_j(X_1, \ldots X_n)$ un estimador de máxima verosimilitud de $\underline{\theta}$. Si $\tau(\underline{\theta}) = (\tau_1(\underline{\theta}), \ldots, \tau_r(\underline{\theta}))$ con $1 \leq r \leq K$ es una transformación de $\underline{\Theta}$ entonces el estimador de máxima verosimilitud de $\tau(\underline{\theta})$ es: $\tau(\hat{\underline{\theta}}) = (\tau_1(\hat{\underline{\theta}}, \ldots, \tau_r(\hat{\underline{\theta}}))$.

Una aplicación de este teorema consiste en estimar las probabilidades marginales de la tabla 4.1.1.; el estimador de máxima verosimilitud de $p_{i.} = \sum_j p_{ij}$ es:

$$\hat{p}_{i.} = \sum_j \hat{p}_{ij} = \sum_j \frac{N_{ij}}{n} = \frac{N_{i.}}{n} \quad \forall \; i = 1, \ldots, I$$

y el estimador de máxima verosimilitud de $p_{\cdot j} = \sum_i p_{ij}$ es:

$$\hat{p}_{\cdot j} = \sum_i \hat{p}_{ij} = \sum_i \frac{N_{ij}}{n} = \frac{N_{\cdot j}}{n} \quad \forall\ j = 1, \ldots, J$$

$\hat{p}_{i\cdot}$ y $\hat{p}_{\cdot j}$ estiman la probabilidad de que un individuo de la población, seleccionado al azar, sea clasificado en A_i y B_j respectivamente.

7. INDEPENDENCIA EN UNA TABLA DE CONTINGENCIA.

7.1 Planteamiento del Problema.

Los datos de la tabla 4.1. fueron recabados con el fin de averiguar si la reacción de los pinos blancos al ser infectados con hongos depende de su edad. Este es un problema que en estadística se puede plantear como una prueba de hipótesis. Las hipótesis a probar son:

H_0: La reacción de los pinos blancos al ser infectados con hongos es independiente de su edad.

H_1: La reacción de los pinos blancos al ser infectados con hongos depende de su edad.

El objetivo es determinar, en base a la muestra, si H_0 es falsa; para ello las hipótesis se reformulan en términos de la distribución conjunta de $N_{11}, \ldots, N_{IJ}$ que, en la tabla 4.1., es una multinomial porque de antemano se determinó el número de pinos a observar.

La hipótesis nula, H_0, señala que la reacción de los pinos

es independiente de su edad; esto quiere decir que, al seleccionar al azar a un pino blanco, la probabilidad de que sea clasificado como sano o enfermo no depende de su clasificación en B.

Simbólicamente esto se expresa como: $p_{ij} = p_{i\cdot} p_{\cdot j} \ \forall i,j$ (la probabilidad de que un pino blanco pertenezca a la celda (A_i, B_j) es igual al producto de las probabilidades marginales de A_i y B_j). Por otra parte, la hipótesis alternativa, H_1 , indica que la reacción de los pinos depende de su edad; es decir, $p_{ij} \neq p_{i\cdot} p_{\cdot j}$ para alguna $i=1,\ldots,I$ y $j=1,\ldots,J$. Por consiguiente, las hipótesis a probar son:

H_0 : Las variables aleatorias $N_{11}, \ldots, N_{IJ}$ se distribuyen como una multinomial con parámetros p_{ij} tales que $p_{ij} = p_{i\cdot} p_{\cdot j} \ \forall i,j$

vs. H_1 : Las variables aleatorias $N_{11}, \ldots, N_{IJ}$ se distribuyen como una multinomial con $p_{ij} \neq p_{i\cdot} p_{\cdot j}$ para alguna $i=1,\ldots,I$ y $j=1,\ldots,J$.

Por brevedad estas hipótesis se enunciarán como sigue: las variables aleatorias $N_{11}, \ldots, N_{IJ}$ se distribuyen como una multinomial con parámetros p_{ij} ; se requiere probar,

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \ \forall i,j \quad \text{vs.} \quad H_1 : p_{ij} \neq p_{i\cdot} p_{\cdot j} \text{ para alguna } i,j \qquad (H\text{-}1)$$

Dos criterios que se han propuesto para determinar si H_0 es falsa son el de la estadística ji-cuadrada y el de la razón de verosimilitudes. El primero de ellos es una prueba de bondad de ajuste porque permite establecer si las observaciones $O_{11}, \ldots, O_{IJ}$ se ajustan o son consistentes con el supuesto de que provienen de una función de densidad multinomial en la que $p_{ij} = p_{i.} p_{.j} \quad \forall i, j$.

7.2 La Estadística ji-Cuadrada.

La idea de la prueba basada en la estadística ji-cuadrada:

$$X^2 = \sum_i \sum_j \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \qquad \text{con} \qquad n\hat{p}_{ij} = \frac{N_{i.} N_{.j}}{n}$$

es comparar, N_{ij}, el número de individuos que pertenecen a la celda (A_i, B_j) con, $\frac{N_{i.} N_{.j}}{n}$, el número de individuos que se espera observar en dicha celda suponiendo que la hipótesis nula es correcta. Se puede demostrar (Everitt, 1977) que la estadística X^2 se distribuye aproximadamente como ji-cuadrada con $k=(I-1)(J-1)$ grados de libertad cuando H_0 es cierta y las N_{ij} no son muy pequeñas. Por esto, el criterio para probar:

$$H_0 : p_{ij} = p_{i.} p_{.j} \ \forall i, j \quad \text{vs.} \quad H_1 : p_{ij} \neq p_{i.} p_{.j} \ \text{para alguna } i, j$$

es: rechazar H_0 si, y solo si, X^2 es mayor que $\chi^{2,(1-\alpha)}_{(k)}$, el cuantil $1-\alpha$ de una ji-cuadrada con $k=(I-1)(J-1)$ grados de libertad. (*)

La estadística ji-cuadrada tiene el defecto de ser directamente proporcional a n como lo señala el teorema 7.2.1.

7.2.1 Teorema.

La estadística X^2 se puede reescribir como:

$$n\left(\sum_i\sum_j \frac{N_{ij}^2}{N_{i\cdot}N_{\cdot j}} - 1\right)$$

Demostración:

Por definición,

$$X^2 = \sum_i\sum_j \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \quad \text{con} \quad n\hat{p}_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{n}$$

Pero,

$$\sum_i\sum_j \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_i\sum_j \left[\frac{N_{ij}^2}{n\hat{p}_{ij}} - \frac{2N_{ij}n\hat{p}_{ij}}{n\hat{p}_{ij}} + \frac{(n\hat{p}_{ij})^2}{n\hat{p}_{ij}}\right]$$

$$= \sum_i\sum_j \frac{N_{ij}^2}{n\hat{p}_{ij}} - 2n + n\sum_i\sum_j \hat{p}_{ij}$$

$$= \sum_i\sum_j \frac{N_{ij}^2}{n\hat{p}_{ij}} - 2n + n\sum_i\sum_j \frac{N_{ij}}{n}$$

$$= \sum_i\sum_j \frac{N_{ij}^2}{n\hat{p}_{ij}} - n = \sum_i\sum_j \frac{N_{ij}^2}{\frac{N_{i\cdot}N_{\cdot j}}{n}} - n = n\left(\sum_i\sum_j \frac{N_{ij}}{N_{i\cdot}N_{\cdot j}} - 1\right)$$

Entonces $X^2 = n\left(\sum_i\sum_j \frac{N_{ij}^2}{N_{i\cdot}N_{\cdot j}} - 1\right)$. □

En virtud del teorema anterior, si n es grande existe el riesgo de rechazar la hipótesis nula aunque ésta sea correcta. Por ejemplo, en la tabla 7.2.1 (Blalock, 196), el valor de la estadística ji-cuadrada es 0.16. Si se elige $\alpha = 0.05$ la hipótesis nula no se rechaza (porque $X^{2,(.95)}_{(1)} = 3.84$). En la -

(*) α es el nivel de la prueba o la probabilidad de cometer el error tipo 1 y se define como la probabilidad de rechazar - la hipótesis nula cuando ésta es correcta. El valor de α se establece antes de llevar a cabo la prueba de hipótesis.

tabla 7.2.2. la proporción de individuos observados en cada celda es

Tabla 7.2.1.

	B_1	B_2	Tot.
A_1	26	24	50
A_2	24	26	50
Tot.	50	50	100

Tabla 7.2.2.

	B_1	B_2	Tot.
A_1	2,600	2,400	5,000
A_2	2,400	2,600	5,000
Tot.	5,000	5,000	10,000

la misma que en la tabla 7.2.1. pero el tamaño de la muestra - es cien veces mayor; por lo tanto, $X^2=16$. En este caso, si - $\alpha=0.05$, la hipótesis de independencia entre las variables se rechaza.

7.3. LA RAZON DE VEROSIMILITUDES.

Las hipótesis (H - 1) también se puede probar usando el - criterio basado en la razón de verosimilitudes generalizada.

7.3.1. Definición: Razón de Verosimilitudes Generalizada.

Sea $L(\theta_1,\ldots,\theta_K;x_1,\ldots,x_n)$ la función de verosimilitud de las variables aleatorias $X_1,\ldots,X_n$ que tienen función de - densidad conjunta $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_K)$ con $\underline{\theta}=(\theta_1,\ldots,\theta_K)\in\underline{\Theta}$. - Para probar las hipótesis:

$$H_o:\underline{\theta}\in\underline{\Theta}_o \quad \text{vs.} \quad H_1:\underline{\theta}\in\underline{\Theta}_o^c \quad \text{con} \quad \underline{\Theta}_o\cup\underline{\Theta}_o^c=\underline{\Theta}$$

se define la razón de verosimilitudes generalizada como:

$$\lambda = \frac{\sup_{\underline{\theta} \in \underline{\Theta}_0} L(\theta_1, \ldots, \theta_k ; x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} L(\theta_1, \ldots, \theta_k ; x_1, \ldots, x_n)}$$

en donde $\underline{\Theta}_0$ es el espacio parametral definido por la hipótesis nula. En la hipótesis (H-1), $\underline{\Theta}_0 = \{ p_{ij} ; p_{ij} = p_{i\cdot} p_{\cdot j} \ \forall \ i,j \}$.

Cuando $H_0 : \underline{\theta} \in \underline{\Theta}_0$ es falsa, el denominador de λ tiende a ser mayor que el numerador. Por esto, la prueba basada en la razón de verosimilitudes generalizada consiste en:

rechazar H_0 si, y solo si, $\lambda \leq \lambda_0$.

λ_0 es una constante fija que toma valores entre cero y uno, su valor depende del nivel de la prueba (α) y de la distribución de Λ (Λ se obtiene al sustituir en λ, las observaciones $x_1, x_2, \ldots, x_n$ por las variables respectivas, $X_1, X_2, \ldots X_n$).

El teorema siguiente indica cuál es la razón de verosimilitudes generalizada correspondiente a la prueba de hipótesis (H-1).

7.3.2. Teorema.

Si la función de densidad conjunta de las variables aleatorias $N_{11}, \ldots, N_{IJ}$ es

$$\ldots \quad f_{N_{11} \ldots N_{IJ}}(o_{11}, \ldots, o_{IJ} ; p_{11}, \ldots, p_{IJ}) = \frac{n!}{\prod_i \prod_j o_{ij}!} \prod_i \prod_j p_{ij}^{o_{ij}}$$

con $\sum_i \sum_j p_{ij} = 1$ y se desea probar:

$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ $\forall$ i,j vs. $H_1 : p_{ij} \neq p_{i\cdot}p_{\cdot j}$ para alguna i,j

entonces, la razón de verosimilitudes generalizada es

$$\hat{\lambda} = \frac{\left(\prod_i o_{i\cdot}^{o_{i\cdot}}\right)\left(\prod_j o_{\cdot j}^{o_{\cdot j}}\right)}{n^n \prod_i \prod_j o_{ij}^{o_{ij}}}.$$

Demostración.

La función de verosimilitud de $N_{11}, \ldots, N_{IJ}$ es

$$L(p_{11}, \ldots, p_{IJ}) = \frac{n!}{\prod_i \prod_j o_{ij}!} \prod_i \prod_j p_{ij}^{o_{ij}} \quad \text{y}$$

$\Theta = \left\{ (p_{11}, \ldots, p_{IJ}) ; \sum_i \sum_j p_{ij} = 1 \right\}$. El espacio parametral definido por H_0 es $\Theta_0 = \left\{ (p_{11}, \ldots, p_{IJ}) ; p_{ij} = p_{i\cdot}p_{\cdot j} \ \forall \ i,j \right\}$. Si $\underline{p} = (p_{11}, \ldots, p_{IJ}) \in \Theta$, el supremo de $L(p_{11}, \ldots, p_{IJ})$ se alcanza en $\hat{\underline{p}} = (\hat{p}_{11}, \ldots, \hat{p}_{IJ})$ con $\hat{p}_{ij} = \frac{o_{ij}}{n}$ $\forall$ i,j. Por lo tanto, $\sup_{\underline{p} \in \Theta} L(p_{11}, \ldots, p_{IJ}) = \frac{n!}{\prod_i \prod_j o_{ij}!} \prod_i \prod_j \left(\frac{o_{ij}}{n}\right)^{o_{ij}}$.

Cuando $\underline{p} \in \Theta_0$, la función de verosimilitud es

$$L(p_{11}, \ldots, p_{IJ}) = \frac{n!}{\prod_i \prod_j o_{ij}!} \prod_i \prod_j (p_{i\cdot}p_{\cdot j})^{o_{ij}}$$

y el supremo de esta función se alcanza en

$$\tilde{p}_{i\cdot} = \frac{o_{i\cdot}}{n} \ \forall \ i=1,\ldots,I \quad \text{y} \quad \tilde{p}_{\cdot j} = \frac{o_{\cdot j}}{n} \ \forall \ j = 1,\ldots,J.$$

Como consecuencia,

$$\sup_{p \in \Theta_0} L(p_{11}, \ldots, p_{IJ}) = L(\tilde{p}_{11}, \ldots, \tilde{p}_{IJ}) = \frac{n!}{\prod_i \prod_j o_{ij}!} \prod_i \prod_j \left(\frac{o_{i.} o_{.j}}{n^2} \right)^{o_{ij}}$$

Entonces,

$$\lambda = \frac{L(\tilde{p}_{11}, \ldots, \tilde{p}_{IJ})}{L(\hat{p}_{11}, \ldots, \hat{p}_{IJ})} = \prod_i \prod_j \left(\frac{o_{i.} o_{.j}}{n\, o_{ij}} \right)^{o_{ij}} = \frac{\left(\prod_i o_{i.}^{o_{i.}} \right) \left(\prod_j o_{.j}^{o_{.j}} \right)}{n^n \prod_i \prod_j o_{ij}^{o_{ij}}}$$

Es decir, $$\lambda = \frac{\left(\prod_i o_{i.}^{o_{i.}} \right) \left(\prod_j o_{.j}^{o_{.j}} \right)}{n^n \prod_i \prod_j o_{ij}^{o_{ij}}}.$$ □

De acuerdo con el criterio basado en la razón de verosimilitudes generalizada, $H_0 : p_{ij} = p_{i.} p_{.j} \ \forall\, i,j$ se rechaza si, y solo si, $\lambda \leq \lambda_0$. En este caso, calcular λ_0 es complicado - (Mood, 1974) porque la distribución de

$$\Lambda = \frac{\left(\prod_i N_{i.}^{N_{i.}} \right) \left(\prod_j N_{.j}^{N_{.j}} \right)}{n^n \prod_i \prod_j N_{ij}^{N_{ij}}}$$

no es única cuando H_0 es correcta (esto es porque en la hipótesis nula no se especifican los valores de $p_{i.}$ y $p_{.j}$ para - toda i,j). Sin embargo, cuando el tamaño de la muestra es - grande, las hipótesis

$$H_0 : p_{ij} = p_{i\cdot}\, p_{\cdot j} \quad \forall\ i,j$$

vs. $H_1 : p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$ para alguna i,j

se pueden probar mediante el criterio siguiente: rechazar H_0 si, y solo si, el valor de $-2 \log \Lambda$ es mayor que $\chi^{2,(1-\alpha)}_{(k)}$, el cuantil $1-\alpha$ de una ji-cuadrada con $k=(I-1)\cdot(J-1)$ grados de libertad. Este criterio (Mood, 1974) está basado en el hecho de que $-2 \log \Lambda$ se distribuye aproximadamente como una ji-cuadrada con $k=(I-1)\cdot(J-1)$ grados de libertad cuando la hipótesis nula es cierta y el tamaño de la muestra es grande.

8. MEDIDAS DE ASOCIACION.

Cabe resaltar que la hipótesis $H_0 : p_{ij} = p_{i\cdot}\, p_{\cdot j}\ \forall\ i,j$ establece que las variables A y B son independientes y H_1 niega este hecho sin especificar algún modelo particular de asociación entre las variables. Por esto, si se rechaza H_0 se sugiere la búsqueda de algún modelo de asociación entre las variables o el empleo de índices que midan dicha asociación.

8.1 Definición: Medida de Asociación.

Una medida de asociación es un índice que resume la información contenida en una tabla de contingencia con el fin de describir la relación que existe entre dos o más variables.

La mayoría de las medidas de asociación han sido diseñadas de tal manera que el valor que tomen esté acotado, superiormente por 1 e inferiormente por 0 ó -1; algunos índices toman valores en $\mathbb{R}^+$ o en $\mathbb{R}$. Cuando las variables son independientes generalmente las medidas de asociación toman el valor de cero y mientras mayor es la relación entre las variables más cercano es su valor a alguna de sus cotas.

8.2. La Asociación Perfecta entre Dos Variables Nominales.

Intuitivamente, dos variables nominales están asociadas en forma perfecta si al saber como fue clasificado un individuo en una de las variables se puede predecir exactamente a qué categoría de la otra variable pertenece.

De acuerdo con Reynolds (1977), la asociación perfecta entre dos variables nominales se puede definir de tres maneras:

8.2.1. Asociación Perfecta Estricta.

Definición: En una tabla de I x I, la asociación entre las variables A y B es perfecta estricta si las categorías de A y B están relacionadas de manera única.

En la tabla 8.2.1., la asociación entre A y B es perfecta estricta porque dicha tabla es de dimensión 3 x 3 y además, cada categoría de A está relacionada con una categoría de B y viceversa (cada B_j está relacionada con una clase de A).

Tabla 8.2.1

	B_1	B_2	B_3
A_1	100	0	0
A_2	0	0	100
A_3	0	100	0

Gráficamente dos variables nominales están relacionadas en forma perfecta estricta si en cada renglón y en cada columna de la tabla solo hay una celda con frecuencia o probabilidad positiva (mayor que cero).

8.2.2. Asociación Perfecta Implícita.

La asociación perfecta implícita, que se define a continuación, es útil cuando la tabla de contingencia es asimétrica (de I X J con $I \neq J$).

Definición: En una tabla de I X J, con $I \neq J$, la asociación entre las variables A y B es perfecta implícita si las categorías de la variable que tiene el mayor número de clases están relacionadas de manera única con las categorías de la otra variable.

La tabla 8.2.2 ilustra este tipo de asociación. En dicha tabla, si se sabe a qué categoría de B pertenece un individuo se puede predecir inequívocamente cuál es su clase en A, pero no la inversa. Esto se debe a que A tiene menos categorías que B.

Tabla 8.2.2.

	B_1	B_2	B_3	B_4
A_1	0	100	0	100
A_2	100	0	0	0
A_3	0	0	100	0

En la tabla anterior, J > I y en cada columna solo hay una celda con frecuencia positiva. Cuando I > J, la asociación entre A y B es perfecta implícita si en cada renglón solo hay una celda con probabilidad o frecuencia positiva.

8.2.3. Asociación Perfecta Débil.

Definición: En una tabla de I x I, la asociación entre las variables A y B es perfecta débil si las frecuencias marginales de A y B son homogéneas y además sus categorías no están relacionadas de manera única.

Cabe señalar que, en esta definición, Reynolds no especifica qué significa que "las frecuencias marginales de dos variables sean homogéneas". Para ejemplificar la asociación perfecta débil Reynolds presenta la tabla siguiente:

Tabla 8.2.3.

	B_1	B_2	B_3	Totales
A_1	100	0	0	0
A_2	100	0	0	100
A_3	100	100	100	300
Totales	300	100	100	500

El valor de una medida de asociación puede variar al usar una u otra definición de asociación perfecta, en parte esto se debe a que las medidas de asociación dependen de la dimensión de la tabla de contingencia.

8.3. Factores que Influyen en el Valor de las Medidas de Asociación.

Hay diversos factores que afectan el valor de las medidas de asociación; algunos de ellos son:

1. El número de categorías consideradas.
2. La existencia de variables relevantes no controladas en el estudio, y
3. La distribución de las frecuencias marginales.

8.3.1. El Número de Categorías Definidas.

Al modificar la dimensión de una tabla de contingencia, el valor de las medidas de asociación se puede alterar. Por ejemplo, en la tabla 8.3.1.1 la asociación entre A y B es perfecta estricta.

Tabla 8.3.1.1.

	B_1	B_2	B_3	B_4
A_1	0	.25	0	0
A_2	.25	0	0	0
A_3	0	0	0	.25
A_4	0	0	.25	0

Si se define una categoría que agrupe a los individuos clasificados en B_2, B_3 y B_4 y otra que agrupe a los individuos pertenecientes a A_2, A_3 y A_4 se obtiene la tabla 8.3.1.2. En esta tabla la asociación entre A y B es perfecta débil.

Tabla 8.3.1.2.

	B_1	B_2'
A_1	0	.25
A_2'	.25	.5

con $B_2' = B_2 \cup B_3 \cup B_4$

$A_2' = A_2 \cup A_3 \cup A_4$

Por lo anterior se recomienda que al hablar del grado de asociación entre dos o más variables nominales se tenga siempre presente cuales fueron las categorías definidas.

8.3.2. Variables no Controladas.

Cuando se estudia el grado de asociación entre dos variables nominales es importante controlar aquellas variables que pudieran atenuar o crear una relación espúrea entre las variables de interés. Este problema se ilustra en el ejemplo siguiente:

8.3.2.1. Ejemplo.

Los datos de la tabla 8.3.2.1.1. (Kendall, 1963) muestran cómo fueron clasificados 450 enfermos que padecían de la misma enfermedad.

Tabla 8.3.2.1.1.

		Recobró la salud		
		Si	No	Totales
Recibió tratamiento médico:	Si	100	200	300
	No	50	100	150
	Totales	150	300	450

En esta tabla el valor de la estadística ji-cuadrada es:

$$X^2 = \frac{(100-100)^2}{100} + \frac{(200-200)^2}{200} + \frac{(50-50)^2}{50} + \frac{(100-100)^2}{100} = 0$$

Por lo tanto, se concluye que el hecho de que un individuo haya recobrado la salud o no, es independiente del tratamiento médico.

Si se realiza el análisis por separado para hombres y mujeres se obtienen las dos tablas siguientes.

Tabla 8.3.2.1.2

<u>Hombres</u>

		Recobró la salud		
		Si	No	Totales
Recibió tratamiento médico:	Si	80	100	180
	No	40	80	120
	Totales	120	180	300

Tabla 8.3.2.1.3

Mujeres

Recibió tratamiento médico:	Recobró la salud: Si	Recobró la salud: No	Totales
Si	20	100	120
No	10	20	30
Totales	30	120	150

En la tabla 8.3.2.1.2, $X^2 = 3.7037$ y en la tabla 8.3.2.1.3, $X^2 = 4.1667$. Por lo tanto, la estadística ji-cuadrada indica - que sí hay una relación entre la recuperación de la salud y el tratamiento médico cuando se hace el análisis para hombres y - mujeres por separado.

Al estudiar el grado de asociación entre dos o más variables se debe hacer todo lo posible por controlar cualquier variable que pudiera influir en dicha asociación.

8.3.3. La Distribución de las Frecuencias Marginales.

Cuando los individuos se concentran en unas cuantas categorías hay sesgo en la distribución de las frecuencias marginales. Esto puede afectar el valor de muchas medidas de asociación y como consecuencia se puede pensar que la relación entre las - variables es más débil de lo que en realidad es. Para evitar - que las medidas de asociación dependan del número de individuos en cada categoría la tabla de contingencia se ajusta de tal manera que las frecuencias marginales de todos los renglones sean

iguales, así como las frecuencias marginales de las columnas. - Una tabla ajustada se conoce como una tabla estandarizada. En - seguida se explica cómo estandarizar tablas de I X J y tablas - de 2 x 2.

8.3.3.1. Estandarización de Tablas de I X J.

En la tabla 8.3.3.1. (Bishop, 1975) se muestra cómo fueron clasificados 500 ingleses de acuerdo a:

A : el país que preferían como aliado en la Guerra Fría y

B : el partido político con el que simpatizaban.

Tabla 8.3.3.1.

Aliado Preferido	Partido Político			
	Derecha	Ninguno	Izquierda	Totales
E.U.A.	225	53	206	484
U.R.S.S.	3	1	12	16
Totales	228	54	218	500

Las frecuencias marginales, sobretodo las de los renglo-- nes, son muy heterogéneas, por esto es conveniente estandarizar la tabla antes de calcular alguna medida de asociación.

El primer paso para ajustar una tabla de contingencia es - especificar qué frecuencias marginales debe tener la tabla es- tandarizada. Si la muestra es de tamaño n y la tabla de contin gencia tiene I renglones, la frecuencia marginal y estandariza- da del i-ésimo renglón, $N_{i.}'$, debe ser $\frac{n}{I}$; además, si la tabla

tiene J columnas, la frecuencia marginal y estandarizada de la j-ésima columna, $N'_{\cdot j}$, debe ser $\frac{n}{J}$. Después, mediante un proceso iterativo se ajustan las frecuencias marginales de A y B. Cada iteración consta de dos pasos. La primera iteración es la siguiente:

Paso 1 : para que la frecuencia marginal del i-ésimo renglón sea igual a $N'_{i\cdot}$, todos los datos del renglón i se multiplican por $\frac{N'_{i\cdot}}{N_{i\cdot}}$. Al efectuar los productos el renglón i queda como sigue:

	B_1		B_j		B_J	Total.
A_i	$N_{i1}^{(1)}$		$N_{ij}^{(1)}$		$N_{iJ}^{(1)}$	$N'_{i\cdot}$

con $N_{ij}^{(1)} = N_{ij}\left(\frac{N'_{i\cdot}}{N_{i\cdot}}\right) \forall\ i=1,..I$

y la tabla de contingencia que se obtiene al ajustar los I renglones es:

	B_1		B_j		B_J	Totales
A_1	$N_{11}^{(1)}$		$N_{1j}^{(1)}$		$N_{1J}^{(1)}$	$N'_{1\cdot}$
A_i	$N_{i1}^{(1)}$		$N_{ij}^{(1)}$		$N_{iJ}^{(1)}$	$N'_{i\cdot}$
A_I	$N_{I1}^{(1)}$		$N_{Ij}^{(1)}$		$N_{IJ}^{(1)}$	$N'_{I\cdot}$
Totales	$N_{\cdot 1}^{(1)}$		$N_{\cdot j}^{(1)}$		$N_{\cdot J}^{(1)}$	n

en donde: $N_{\cdot j}^{(1)} = \sum_i N_{ij}^{(1)} \quad \forall \quad j=1,..,J$

Paso 2 : en seguida se ajustan las frecuencias marginales de B; para ello la columna j se multiplica por $\frac{N'_{\cdot j}}{N^{(1)}_{\cdot j}}$.

La j-ésima columna estandarizada es:

	B_j
A_1	$N_{1j}^{(2)}$
A_i	$N_{ij}^{(2)}$
A_I	$N_{Ij}^{(2)}$
Total	$N'_{\cdot j}$

con $N_{ij}^{(2)} = N_{ij}^{(1)} \left(\frac{N'_{\cdot j}}{N^{(1)}_{\cdot j}} \right) \quad \forall \; j = 1, .., J.$

y las J columnas ajustadas son:

	B_1		B_j		B_J	Totales
A_1	$N_{11}^{(2)}$		$N_{1j}^{(2)}$		$N_{1J}^{(2)}$	$N_{1\cdot}^{(2)}$
A_i	$N_{i1}^{(2)}$		$N_{ij}^{(2)}$		$N_{iJ}^{(2)}$	$N_{i\cdot}^{(2)}$
A_I	$N_{I1}^{(2)}$		$N_{Ij}^{(2)}$		$N_{IJ}^{(2)}$	$N_{I\cdot}^{(2)}$
Totales	$N'_{\cdot 1}$		$N'_{\cdot j}$		$N'_{\cdot J}$	n

en donde: $N_{i\cdot}^{(2)} = \sum_j N_{ij}^{(2)} \quad \forall \quad i = 1, .., I$

La primera iteración termina al estandarizar las categorías de B. Pero, al hacerlo, las frecuencias marginales de A

ya no son iguales a $N'_{i.}$; por lo tanto, es necesario volver a - ajustarlas iniciándose de esta manera una nueva iteración. El - proceso finaliza cuando la magnitud de las diferencias entre - las frecuencias N_{ij} de iteraciones consecutivas sea menor que - una constante positiva fijada de antemano.

En la tabla 8.3.3.1, n=500, I=2 y J=3 por lo que las - frecuencias marginales de la tabla estandarizada deben ser:

$$N'_{i.} = \frac{500}{2} = 250 \quad \forall\, i = 1,2$$

$$\text{y} \quad N'_{.j} = \frac{500}{3} = 166.6 \quad \forall\, j = 1,2,3$$

Iteración 1.

Paso 1 : para ajustar los totales de los renglones se multiplica el primero de ellos por $\frac{N'_{1.}}{N_{1.}} = \frac{250}{484} = 0.516$ y el segundo por $\frac{N'_{2.}}{N_{2.}} = \frac{250}{16} = 15.625$. La tabla - resultante es:

	Derecha	Ninguno	Izquierda	Totales
E.U.A.	116.219	27.376	106.4049	250
U.R.S.S.	46.875	15.625	187.5	250
Tot.	163.094	42.9626	293.9094	500

Paso 2 : las frecuencias marginales de esta última tabla - se ajustan multiplicando la columna 1 por $\frac{N.'_1}{N.^{(1)}_1} = \frac{166.6}{163.094}$, la columna 2 por $\frac{N.''_2}{N.^{(1)}_2} = \frac{166.6}{42.9626}$ y la - columna 3 por $\frac{N.'_3}{N.^{(1)}_3} = \frac{166.6}{293.9049}$. La tabla que se obtiene es:

	Derecha	Ninguno	Izquierda	Totales
E.U.A.	118.7648	106.20089	60.3397	285.305
U.R.S.S.	47.9018	60.61473	106.3269	214.8439
Totales	166.66	166.8156	166.66	500

En esta tabla las frecuencias marginales de los renglones son distintas a 250; sin embargo, las frecuencias que se obtuvieron no son tan heterogéneas como en la tabla 8.3.3.1. Los - totales de los renglones se pueden volver a ajustar en otra - iteración.

8.3.3.2. Estandarización de Tablas de 2 X 2.

Las tablas de contingencia de 2 X 2 se pueden estandarizar usando el procedimiento descrito en la sección anterior o usando la transformación propuesta en el teorema 8.3.3.2.1. Este - teorema se enuncia en términos de la tabla de contingencia para la población y, en Capítulos posteriores, permitirá demostrar - igualdades entre varias medidas de asociación.

8.3.3.2.1. Teorema.

Se tiene una tabla de contingencia de 2 X 2 como la siguiente:

	B_1	B_2	Totales
A_1	p_{11}	p_{12}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	$p_{2\cdot}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	1

con $p_{11} \neq 0$, $p_{12} \neq 0$ y $p_{21} \neq 0$

Si la transformación $p_{ij} \longrightarrow r_i c_j p_{ij} = p'_{ij}$ es tal que: $p'_{i\cdot} = p'_{\cdot j} = \frac{1}{2}$ $\forall$ $i=1,2$ y $j = 1, 2$ entonces:

$$p'_{11} = p'_{22} = \frac{\sqrt{\alpha}}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right)$$

y

$$p'_{12} = p'_{21} = \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right)$$

con $\alpha = \dfrac{p_{11}\, p_{22}}{p_{12}\, p_{21}}$

Demostración.

El objetivo es encontrar p'_{11}, p'_{12}, p'_{21} y p'_{22} tales que:

$$p'_{11} + p'_{12} = p'_{21} + p'_{22} = p'_{11} + p'_{21} = p'_{12} + p'_{22} = 0.5$$

Como por hipótesis $p'_{ij} = r_i c_j p_{ij}$ $\forall$ $i = 1, 2$ y $j = 1,2$ entonces, las igualdades anteriores son equivalentes a:

$$r_1 c_1 p_{11} + r_1 c_2 p_{12} = 0.5 \qquad \ldots (1)$$

$$r_2 c_1 p_{21} + r_2 c_2 p_{22} = 0.5 \qquad \ldots (2)$$

$$r_1 c_1 p_{11} + r_2 c_1 p_{21} = 0.5 \qquad \ldots (3)$$

$$r_1 c_2 p_{12} + r_2 c_2 p_{22} = 0.5 \qquad \ldots (4)$$

Estas cuatro ecuaciones son linealmente dependientes; por lo tanto, hay una infinidad de soluciones para r_1, r_2, c_1 y c_2. Una de estas soluciones se obtiene como sigue:

Si se factoriza r_1 en la ecuación (1), r_2 en (2) y c_2 en (4) se obtiene que:

$$r_1 (c_1 p_{11} + c_2 p_{12}) = 0.5 \Rightarrow r_1 = \frac{1}{2 (c_1 p_{11} + c_2 p_{12})} \ldots (1)'$$

$$r_2 (c_1 p_{21} + c_2 p_{22}) = 0.5 \Rightarrow r_2 = \frac{1}{2 (c_1 p_{21} + c_2 p_{22})} \ldots (2)'$$

$$c_2 (r_1 p_{12} + r_2 p_{22}) = 0.5 \Rightarrow c_2 = \frac{1}{2 (r_1 p_{12} + r_2 p_{22})} \ldots (3)'$$

Si se sustituyen (1)' y (2)' en (4) :

$$c_2 = \frac{1}{2 \left[\frac{p_{12}}{2 (c_1 p_{11} + c_2 p_{12})} + \frac{p_{22}}{2 (c_1 p_{21} + c_2 p_{22})} \right]}$$

$$= \frac{(c_1 p_{11} + c_2 p_{12})(c_1 p_{21} + c_2 p_{22})}{p_{12} (c_1 p_{21} + c_2 p_{22}) + p_{22} (c_1 p_{11} + c_2 p_{12})}$$

Esto implica que:

$$c_2 (c_1 p_{12} p_{21} + c_2 p_{12} p_{22} + c_1 p_{11} p_{22} + c_2 p_{12} p_{22}) =$$

$$c_1^2 p_{11} p_{21} + c_1 c_2 p_{11} p_{22} + c_1 c_2 p_{12} p_{21} + c_2^2 p_{12} p_{22}$$

$$\Leftrightarrow \quad c_2^2 p_{12} p_{22} = c_1^2 p_{11} p_{21} \quad \Leftrightarrow \quad c_1 = c_2 \sqrt{\frac{p_{12} p_{22}}{p_{11} p_{21}}}$$

Si $c_2 = \sqrt{p_{11} p_{21}}$ entonces $c_1 = \sqrt{p_{12} p_{22}}$

Si se sustituyen c_1 y c_2 en (1)' se obtiene que:

$$r_1 = \frac{1}{2\left(p_{11}\sqrt{p_{12} p_{22}} + p_{12}\sqrt{p_{11} p_{21}}\right)} = \frac{1}{2\sqrt{p_{11} p_{12}}\left(\sqrt{p_{11} p_{22}} + \sqrt{p_{12} p_{21}}\right)}$$

y se sustituyen c_1 y c_2 en (2)'

$$r_2 = \frac{1}{2\left(p_{21}\sqrt{p_{12} p_{22}} + p_{22}\sqrt{p_{11} p_{21}}\right)} = \frac{1}{2\sqrt{p_{12} p_{22}}\left(\sqrt{p_{12} p_{21}} + \sqrt{p_{11} p_{22}}\right)}$$

Por lo tanto:

$$p_{11}' = r_1 c_1 p_{11} = \frac{p_{11}}{2}\sqrt{\frac{p_{12} p_{22}}{p_{11} p_{12}}}\left(\frac{1}{\sqrt{p_{11} p_{22}} + \sqrt{p_{12} p_{21}}}\right) = \frac{\sqrt{p_{11} p_{22}}}{2} \cdot$$

$$\left(\frac{1}{\sqrt{p_{11} p_{22}} + \sqrt{p_{12} p_{21}}}\right) = \frac{1}{2}\sqrt{\frac{p_{11} p_{22}}{p_{12} p_{21}}}\left(\frac{1}{\sqrt{\frac{p_{11} p_{22}}{p_{12} p_{21}}} + 1}\right) =$$

$$\frac{\sqrt{\alpha}}{2} \cdot \left(\frac{1}{\sqrt{\alpha} + 1}\right) \quad \text{con} \quad \alpha = \frac{p_{11} p_{22}}{p_{12} p_{21}}$$

Análogamente:

$$p''_{22} = r_2 c_2 p_{22} = \frac{\sqrt{\alpha}}{2} \cdot \left(\frac{\quad}{\sqrt{\alpha \quad - \quad 1}} \right)$$

$$p'_{12} = r_1 c_2 p_{12} = \frac{1}{2} \cdot \left(\frac{\quad}{\sqrt{\quad \quad + 1}} \right)$$

y $$p'_{21} = r_2 c_1 p_{21} = \frac{1}{2} \cdot \left(\frac{\quad}{\sqrt{\quad \quad + 1}} \right)$$

Entonces:

$$p'_{11} = p'_{22} = \frac{\sqrt{\alpha}}{2} \cdot \left(\frac{\quad}{\sqrt{\alpha \quad - \quad 1}} \right)$$

y $$p'_{12} = p'_{21} = \frac{1}{2} \cdot \left(\frac{\quad}{\sqrt{\alpha \quad - \quad 1}} \right)$$

$$\text{con } \alpha = \frac{p_{11} p_{22}}{p_{21} p_{12}}$$

□

8.3.3.2.2. Ejemplo:

Con el propósito de ilustrar cómo se aplica el teorema anterior, considérese la siguiente tabla de 2 × 2:

	B_1	B_2	Totales
A_1	.25	.5	.75
A_2	.05	.2	.25
Totales	.3	.7	1

Análogamente:

$$p'_{22} = r_2 c_2 p_{22} = \frac{\sqrt{\alpha}}{2} \cdot \left(\frac{1}{\sqrt{\alpha} + 1} \right)$$

$$p'_{12} = r_1 c_2 p_{12} = \frac{1}{2} \cdot \left(\frac{1}{\sqrt{\alpha} + 1} \right)$$

y $$p'_{21} = r_2 c_1 p_{21} = \frac{1}{2} \cdot \left(\frac{1}{\sqrt{\alpha} + 1} \right)$$

Entonces:

$$p'_{11} = p'_{22} = \frac{\sqrt{\alpha}}{2} \cdot \left(\frac{1}{\sqrt{\alpha} + 1} \right)$$

y $$p'_{12} = p'_{21} = \frac{1}{2} \cdot \left(\frac{1}{\sqrt{\alpha} + 1} \right)$$

$$\text{con} \quad \alpha = \frac{p_{11} p_{22}}{p_{21} p_{12}}$$

□

8.3.3.2.2. Ejemplo:

Con el propósito de ilustrar cómo se aplica el teorema anterior, considérese la siguiente tabla de 2 X 2:

	B_1	B_2	Totales
A_1	.25	.5	.75
A_2	.05	.2	.25
Totales	.3	.7	1

Para estandarizar esta tabla solo es necesario calcular - α, p'_{11} y p'_{12}.

Como $$\alpha = \frac{p_{11}\, p_{22}}{p_{12}\, p_{21}} = \frac{0.25\ (0.20)}{0.50\ (0.05)} = 2$$

entonces $$p'_{11} = p'_{22} = \frac{\sqrt{\alpha}}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right) = \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}+1}\right) \simeq 0.2928$$

y $$p'_{12} = p'_{21} = \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right) = \frac{1}{2}\left(\frac{1}{\sqrt{2}+1}\right) \simeq 0.2071$$

Por lo tanto, la tabla estandarizada es :

	B_1	B_2	Totales
A_1	.293	.207	.5
A_2	.207	.293	.5
Totales	.5	.5	1

9. El Coeficiente de Correlación de Pearson.

El coeficiente de correlación de Pearson es un índice muy usado para medir el grado de asociación lineal entre dos variables cuantitativas. Este coeficiente ha sido tomado como modelo para diseñar algunas medidas de asociación para variables - nominales.

9.1 Definición.

El coeficiente de correlación poblacional entre las variables X y Y, ρ_{xy} , se define como:

$$\rho_{xy} = \frac{\text{cov}\ (X,Y)}{\sigma_x\ \sigma_y}$$

en donde: $\text{cov}\ (X,Y) = E\left[(X - \mu_x)\cdot(Y - \mu_y)\right] = E\ (XY) - \mu_x\mu_y$

μ_x y μ_y son, respectivamente, la media de X y Y.

y σ_x y σ_y son, respectivamente, la desviación estándar de X y Y; ambas deben ser mayores que cero.

9.2 Propiedades.

9.2.1. ρ_{xy} mide la relación lineal que existe entre X y Y en el sentido de que ρ_{xy} es positivo cuando $X - \mu_x$ y $Y - \mu_y$ tienden a ser mayores o menores que cero y ρ_{xy} es negativo cuando $X - \mu_x$ y $Y - \mu_y$ tienden a tener signos opuestos.

9.2.2. El coeficiente de correlación toma valores en el intervalo $[-1,1]$.

9.2.3. $\rho_{xy} = 1$ si la asociación lineal entre X y Y es perfecta y positiva, es decir, si a incrementos o decrementos de "b" unidades en una de las variables le corresponden respectivamente incrementos o decrementos de "c" unidades en la otra variable.

9.2.4. $\rho_{xy} = -1$ si la asociación lineal entre X y Y es perfecta y negativa; es decir, si a incrementos de "b" unidades en una de las variables le corresponden decrementos de "c" unidades en la otra.

9.2.5. $\rho_{xy}=0$ si X y Y son independientes.

9.2.6. El coeficiente de correlación es adimensional (no depende de las unidades de X y Y).

9.3. Interpretación. (*)

Cuando X y Y son, respectivamente, la variable independiente y la variable dependiente, el coeficiente de correlación, elevado al cuadrado, ρ_{xy}^2, se interpreta como la proporción de la variación de Y que queda explicada por la variación que tiene X. A ρ_{xy}^2 se le conoce como el coeficiente de determinación.

9.4. Estimador de Máxima Verosimilitud.

Si $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una muestra de tamaño n de las variables X y Y, el estimador de máxima verosimilitud de ρ_{xy} es:

$$\hat{\rho}_{xy} = r_{xy} = \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})\ (Y_i-\bar{Y})}{\sqrt{\frac{\sum_i (X_i-\bar{X})^2}{n-1}}\sqrt{\frac{\sum_i (Y_i-\bar{Y})^2}{n-1}}} = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_i (X_i-\bar{X})^2 \sum_i (Y_i-\bar{Y})^2}}$$

A r_{xy} se le conoce como el coeficiente de correlación muestral.

(*) Esta interpretación solo es válida en el contexto de análisis de regresión.

9.5 Ejemplo.

La tabla 9.5. (Infante y Zárate, 1986) muestra cuál fue el peso (X) y la estatura (Y) de 10 personas.

Tabla 9.5.

Individuo	1	2	3	4	5	6	7	8	9	10
X : Peso (en kg)	63	52	78	49	71	62	68	48	56	67
Y : Estatura (en cm)	162	158	167	151	162	168	167	153	152	173

Como $n = 10$ $\bar{x} = 61.4$, $\bar{y} = 161.3$; $\sum_i x_i y_i = 99.568$,

$$\sum_i (x_i - \bar{x})^2 = 896.4 \quad y \quad \sum_i (y_i - \bar{y})^2 = 520.10$$

entonces: $r_{xy} = \dfrac{529.8}{682.80} = 0.77592$

En este ejemplo r_{xy}^2 no tiene una interpretación porque al realizar el estudio no se supuso que existía alguna relación de dependencia entre las variables.

10. BIBLIOGRAFIA.

1. Birch, M.W.:
Maximum Likelihood in Three-way Contingency Tables.
Journal of the Royal Statistical Society,
1963; Series B, 25 : 220-233.

2. Bishop, Y.M.M., Fienberg, S., Holland, P.W.:
Discrete Multivariate Analysis: Theory and Practice.
MIT, Cambridge Mass., 1975.

3. Blalock, H.M.:
Social Statistics.
McGraw-Hill, New York, 1960.

4. Conover, W.L.:
Practical Nonparametric Statistics.
John Wiley, New York, 2nd. Ed., 1980.

5. Everitt, B.S.:
The Analysis of Contingency Tables.
Chapman and Hall, London, 1977.

6. Fienberg, S.E.:
The Analysis of Cross-Classified Categorical Data.
MIT Press, Cambridge, Massachusetts and London, England, 1977.

7. Goodman, L.A., Kruskal, W.H.:
Measures of association for cross classifications.
Journal of the American Statistical Association, 1954; 49: 732 - 763.

8. Infante, S.G., Zárate, de L.G.P.:
Métodos Estadísticos (Un Enfoque Interdisciplinario).
Ed. Trillas, S.A. de C.V.,
México, Argentina, España, Colombia, Puerto Rico, Venezuela, 2a. Ed., 1984.

9. Kendall, M.G., Stuart, A.:
The Advanced Theory of Statistics.
C. Griffin and Co., Ltd., London, 2nd. Ed., Vol. 2: Inference and Relationship, 1963.

10. Méndez, I.R.:
Estadística y Método Científico.
Comunicaciones Técnicas, Vol. 2, Serie A: General, No. 13.
Ed. Centro de Investigación en Matemáticas Aplicadas y en Sistemas, UNAM, 1975.

11. Mood, A.M., Graybill, F.A., Boes, D.C.:
Introduction to the Theory of Statistics.
Mc Graw-Hill International Book Co., Tokyo, 3rd. Ed., 1963.

12. Reynolds, H.T.:
The Analysis of Cross-Classifications.
The Free Press, New York, 1977.

13. Steel, R.G.D., Torrie, J.H.:
Principles and Procedures of Statistics (A Biometrical Approach).
McGraw-Hill International Book Co., Singapore, 2nd. Ed., 1980

II. MEDIDAS BASADAS EN LA ESTADISTICA JI - CUADRADA

1. El coeficiente de contingencia en media cuadrática.

2. El coeficiente de contingencia de Karl Pearson.

3. El índice de Tschuprov.

4. La V de Cramér.

5. El coeficiente de correlación.

6. Ejemplo.

7. Bibliografía.

II. MEDIDAS BASADAS EN LA ESTADISTICA JI-CUADRADA

En este Capítulo se menciona un grupo de medidas que son - función de la estadística X^2:

1. El coeficiente de contingencia en media cuadrática,
2. El coeficiente de contingencia de Karl Pearson,
3. El índice de Tschuprov, y
4. La V de Cramér.

Todas estas medidas se pueden interpretar cuando las variables son independientes o cuando están asociadas en forma perfecta estricta. El coeficiente de contingencia en media cuadrática, además, se puede interpretar cuando las dos variables en estudio son dicotómicas y asimétricas.

Otra medida de asociación que se considera en este Capítulo es el coeficiente de correlación entre dos variables nominales y dicotómicas. Este índice se obtiene codificando las dos categorías de las variables en estudio y usando la definición - del coeficiente de correlación de Pearson.

Para cada medida se incluye, sin demostración, la expresión de la varianza asintótica del índice muestral. Dicha varianza permite construir intervalos de confianza aproximados - (válidos únicamente cuando el tamaño de la muestra es grande) - para el índice poblacional.

1. EL COEFICIENTE DE CONTINGENCIA EN MEDIA CUADRATICA.

1.1 Definición y Estimador.

De las cuatro medidas de asociación basadas en la estadística ji-cuadrada que se mencionaron inicialmente, el coeficiente de contingencia en media cuadrática es el más antiguo. El índice poblacional se define como:

$$\Phi^2 = \sum_i \sum_j \frac{(p_{ij} - p_{i\cdot}.p_{\cdot j})^2}{p_{i\cdot}.p_{\cdot j}} = \sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}.p_{\cdot j}} - 1 \quad \ldots \text{(E-1.1.)}$$

y su estimador de máxima verosimilitud (Bishop, 1975; Reynolds, 1977) se obtiene dividiendo la estadística X^2 por n, es decir: $\hat{\Phi}^2 = \frac{X^2}{n}$.

1.2 Propiedades.

1.2.1. Φ^2, a diferencia de la estadística X^2, no depende del tamaño de la muestra.

1.2.2. Φ^2 toma valores en el intervalo $[0,1]$ si la tabla es de 2 X 2 y toma valores en el intervalo $[0,\infty)$ si la tabla es mayor (Bishop, 1975).

1.2.3. Cuando las variables son independientes, $\Phi^2 = 0$ (porque $p_{ij} = p_{i\cdot}.p_{\cdot j}$).

1.3 Desventajas.

1.3.1. El valor del coeficiente de contingencia en media cuadrática depende de las frecuencias (o probabilidades) marginales. Como consecuencia, los valores de Φ^2 no son comparables cuando se tienen tablas de contingencia con frecuencias marginales distintas.

1.3.2 Φ^2 solo se puede interpretar cuando las variables son independientes o cuando están asociadas en forma perfecta estricta o perfecta implícita. (*)

Los dos teoremas siguientes indican qué valor toma Φ^2 cuando las variables están asociadas en forma perfecta estricta o perfecta implícita. Unicamente se demuestra el primer teorema; para el segundo se dá un ejemplo.

1.4 Teorema.

Si la tabla de contingencia es de I X I y la asociación entre las variables A y B es perfecta estricta, el coeficiente de contingencia en media cuadrática es igual a I-1.

Demostración:

Como por hipótesis, la asociación entre las variables es perfecta estricta entonces en cada renglón y en cada columna de la tabla solo hay una p_{ii} distinta de cero. A esta p_{ii} se le denotará como p^*_{ii}. Además $p_{i\cdot} = p_{\cdot i} = p^*_{ii}$ para toda $i = 1, \ldots, I$ por lo tanto:

$$\Phi^2 = \sum_i \frac{(p^*_{ii})^2}{(p^*_{ii})(p^*_{ii})} - 1 = I - 1 \qquad \ldots \text{ (E-1.4.)}$$

□

(*) Posteriormente se demostrará que si la tabla es de 2 X 2 y las variables son asimétricas, cualquier valor que tome Φ^2 se puede interpretar.

1.5 Teorema.

Si la tabla de contingencia es de I x J y las variables están asociadas en forma perfecta implícita entonces $\Phi^2 = \text{mín}\{I-1, J-1\}$.

La tabla siguiente tiene más columnas que renglones y en cada columna solo hay una celda cuya probabilidad es distinta de cero. Por lo tanto A y B. están asociados en forma perfecta implícita.

	B_1	B_2	B_3	B_4	Totales
A_1	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$
A_2	0	$\frac{1}{6}$	$\frac{1}{4}$	0	$\frac{5}{12}$
A_3	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$
Totales	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{4}$	1

Al calcular el coeficiente de contingencia en media cuadrática usando la expresión (E-1.1.) se tiene que:

$$\Phi^2 = \frac{p_{11}^2}{p_{1\cdot}p_{\cdot 1}} + \frac{p_{22}^2}{p_{2\cdot}p_{\cdot 2}} + \frac{p_{23}^2}{p_{2\cdot}p_{\cdot 3}} + \frac{p_{34}^2}{p_{3\cdot}p_{\cdot 4}} - 1$$

$$= 1 + \frac{2}{5} + \frac{3}{5} + 1 - 1 = 2$$

Es decir, $\Phi^2 = 2 = \text{mín}\{3-1, 4-1\}$ como lo indica el teorema 1.5.

El interés del teorema que se enuncia en seguida radica en que el coeficiente de contingencia en media cuadrática puede ser expresado de diversas maneras cuando la tabla de contingencia es de I X 2.

1.6 Teorema.

Cuando la tabla de contingencia es de I X 2, el coeficiente de contingencia en media cuadrática se puede enunciar en las siguientes formas:

$$\sum_{i=1}^{I}\sum_{j=1}^{2}\left(\frac{p_{ij}}{p_{\cdot j}}-p_{i\cdot}\right)^2\frac{p_{\cdot j}}{p_{i\cdot}}$$

$$\sum_{i=1}^{I}\left(\frac{p_{i1}}{p_{\cdot 1}}-\frac{p_{i2}}{p_{\cdot 2}}\right)^2\frac{p_{\cdot 1}p_{\cdot 2}}{p_{i\cdot}}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{2}\left(\frac{p_{ij}}{p_{i\cdot}}-p_{\cdot j}\right)^2\frac{p_{i\cdot}}{p_{\cdot j}}$$

$$\sum_{i=1}^{I}\left(\frac{p_{i1}}{p_{i\cdot}}-p_{\cdot 1}\right)^2\frac{p_{i\cdot}}{p_{\cdot 1}p_{\cdot 2}}$$

$$1-\sum_{i=1}^{I}\frac{p_{i1}p_{i2}}{p_{\cdot 1}p_{\cdot 2}p_{i\cdot}}$$

Demostraciones

1.6.1. Primero se demuestra que $\Phi^2=\sum_{i=1}^{I}\sum_{j=1}^{2}\left(\frac{p_{ij}}{p_{\cdot j}}-p_{i\cdot}\right)^2\frac{p_{\cdot j}}{p_{i\cdot}}$

Por definición, $\Phi^2=\sum_{i=1}^{I}\sum_{j=1}^{2}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$

y $\sum_{i=1}^{I}\sum_{j=1}^{2}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}=\sum_i\sum_j\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2p_{\cdot j}}{p_{\cdot j}^2p_{i\cdot}}=\sum_i\sum_j\left(\frac{p_{ij}}{p_{\cdot j}}-p_{i\cdot}\right)^2\frac{p_{\cdot j}}{p_{i\cdot}}$

Entonces $\Phi^2=\sum_{i=1}^{I}\sum_{j=1}^{2}\left(\frac{p_{ij}}{p_{\cdot j}}-p_{i\cdot}\right)^2\frac{p_{\cdot j}}{p_{i\cdot}}$

1.6.2. A continuación se demuestra que $\Phi^2=\sum_{i=1}^{I}\left(\frac{p_{i1}}{p_{\cdot 1}}-\frac{p_{i2}}{p_{\cdot 2}}\right)^2\frac{p_{\cdot 1}p_{\cdot 2}}{p_{i\cdot}}$

$$\Phi^2=\sum_{i=1}^{I}\sum_{j=1}^{2}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}=\sum_i\left[\frac{(p_{i1}-p_{i\cdot}p_{\cdot 1})^2}{p_{i\cdot}p_{\cdot 1}}+\frac{(p_{i2}-p_{i\cdot}p_{\cdot 2})^2}{p_{i\cdot}p_{\cdot 2}}\right]$$

$$= \sum_i \left[\frac{p_{.2}(p_{i1} - p_{i.}p_{.1})^2 + p_{.1}(p_{i2} - p_{i.}p_{.2})^2}{p_{i.}p_{.1}p_{.2}} \right]$$

en donde:

$$p_{.2}(p_{i1} - p_{i.}p_{.1})^2 + p_{.1}(p_{i2} - p_{i.}p_{.2})^2 = p_{.2}(p_{i1} - p_{i.}p_{.1})^2 + (1 - p_{.2})(p_{i2} - p_{i.}p_{.2})^2$$

$$= p_{.2}\left[(p_{i1} - p_{i.}p_{.1})^2 - (p_{i2} - p_{i.}p_{.2})^2\right] + (p_{i2} - p_{i.}p_{.2})^2$$

$$= p_{.2}\left[(p_{i1} - p_{i.}p_{.1}) - (p_{i2} - p_{i.}p_{.2})\right]\left[(p_{i1} - p_{i.}p_{.1}) + (p_{i2} - p_{i.}p_{.2})\right] + (p_{i2} - p_{i.}p_{.2})^2$$

$$= p_{.2}\left[(p_{i1} - p_{i.}p_{.1}) - (p_{i2} - p_{i.}p_{.2})\right]\left[(p_{i1} + p_{i2}) - p_{i.}(p_{.1} + p_{.2})\right] + (p_{i2} - p_{i.}p_{.2})^2$$

$$= p_{.2}\left[(p_{i1} - p_{i.}p_{.1}) - (p_{i2} - p_{i.}p_{.2})\right](p_{i.} - p_{i.}) + (p_{i2} - p_{i.}p_{.2})^2 = (p_{i2} - p_{i.}p_{.2})^2$$

entonces:

$$\sum_i \left[\frac{p_{.2}(p_{i1} - p_{i.}p_{.1})^2 + p_{.1}(p_{i2} - p_{i.}p_{.2})^2}{p_{i.}p_{.1}p_{.2}} \right] = \sum_i \frac{(p_{i2} - p_{i.}p_{.2})^2}{p_{i.}p_{.1}p_{.2}}$$

$$= \sum_i \left[\frac{p_{.2}(p_{i1} + p_{i2}) - p_{i2}}{p_{i.}p_{.1}p_{.2}} \right]^2 = \sum_i \left[\frac{p_{i1}p_{.2} - p_{i2}(p_{.2} + 1)}{p_{i.}p_{.1}p_{.2}} \right]^2$$

$$= \sum_i \frac{(p_{i1}p_{.2} - p_{i2}p_{.1})^2}{p_{i.}p_{.1}p_{.2}} = \sum \frac{(p_{i1}p_{.2} - p_{i2}p_{.1})^2 p_{.1}p_{.2}}{p_{i.}(p_{.1}p_{.2})^2}$$

$$= \sum_i \left(\frac{p_{i1}p_{.2}}{p_{.1}p_{.2}} - \frac{p_{i2}p_{.1}}{p_{.1}p_{.2}} \right)^2 \frac{p_{.1}p_{.2}}{p_{i.}} = \sum_i \left(\frac{p_{i1}}{p_{.1}} - \frac{p_{i2}}{p_{.2}} \right)^2 \frac{p_{.1}p_{.2}}{p_{i.}}$$

Por lo tanto, $\Phi^2 = \sum_i \left(\frac{p_{i1}}{p_{.1}} - \frac{p_{i2}}{p_{.2}} \right)^2 \frac{p_{.1}p_{.2}}{p_{i.}}$

1.6.3. Los pasos que deben seguirse para comprobar que $\Phi^2 = \sum_{i=1}^{I} \sum_{j=1}^{2} \left(\frac{p_{ij}}{p_{i.}} - p_{.j} \right)^2 \frac{p_{i.}}{p_{.j}}$ son análogos a los de la demostración 1.6.1.

1.6.4. En seguida se demuestra que

$$\Phi^2 = \sum_i \left(\frac{p_{i1}}{p_{i.}} - p_{.1} \right)^2 \frac{p_{i.}}{p_{.1}p_{.2}}$$

De acuerdo con el inciso 1.6.2.:

$$\Phi^2 = \sum_i \left(\frac{p_{i1}}{p_{.1}} - \frac{p_{i2}}{p_{.2}} \right)^2 \frac{p_{.1}\, p_{.2}}{p_{i.}} \quad ; \quad \text{además}$$

$$\sum_i \left(\frac{p_{i1}}{p_{.1}} - \frac{p_{i2}}{p_{.2}} \right)^2 \frac{p_{.1}\, p_{.2}}{p_{i.}} = \sum_i \left(\frac{p_{i1}\, p_{.2}}{p_{.1}\, p_{.2}} - \frac{p_{i2}\, p_{.1}}{p_{.1}\, p_{.2}} \right)^2 \frac{p_{.1}\, p_{.2}}{p_{i.}}$$

$$= \sum_i \frac{(p_{i1}\, p_{.2} - p_{i2}\, p_{.1})^2}{(p_{.1}\, p_{.2})^2} \frac{p_{.1}\, p_{.2}}{p_{i.}} = \sum_i \frac{(p_{i1}\, p_{.2} - p_{i2}\, p_{.1})^2}{p_{i.}\, p_{.1}\, p_{.2}}$$

$$= \sum_i \frac{\left[p_{i1}(1 - p_{.1}) - p_{i2}\, p_{.1} \right]^2}{p_{i.}\, p_{.1}\, p_{.2}} = \sum_i \frac{\left[p_{i1} - (p_{i1} + p_{i2})\, p_{.1} \right]^2}{p_{i.}\, p_{.1}\, p_{.2}}$$

$$= \sum_i \left(\frac{p_{i1} - p_{i.}\, p_{.1}}{p_{i.}} \right)^2 \frac{p_{i.}}{p_{.1}\, p_{.2}} = \sum_i \left(\frac{p_{i1}}{p_{i.}} - p_{.1} \right)^2 \frac{p_{i.}}{p_{.1}\, p_{.2}}$$

Entonces $$\Phi^2 = \sum_i \left(\frac{p_{i1}}{p_{i.}} - p_{.1} \right)^2 \frac{p_{i.}}{p_{.1}\, p_{.2}}$$

1.6.5. Por último se demuestra que $$\Phi^2 = 1 - \sum_i \frac{p_{i1}\, p_{i2}}{p_{.1}\, p_{.2}\, p_{i.}}$$

De acuerdo con la expresión (E - 1.1.)

$$\Phi^2 = \sum_i \sum_j \frac{p_{ij}^2}{p_{i.}\, p_{.j}} - 1$$

Como

$$\sum_{i=1}^{I} \sum_{j=1}^{2} \frac{p_{ij}^2}{p_{i.}\, p_{.j}} - 1 = \sum_i \left(\frac{p_{i1}^2}{p_{i.}\, p_{.1}} + \frac{p_{i2}^2}{p_{i.}\, p_{.2}} \right) - 1 = \sum_i \left(\frac{p_{i1}}{p_{i.}} \right) \left(\frac{p_{i1}}{p_{.1}} \right) + \sum_i \left(\frac{p_{i2}}{p_{i.}} \right) \left(\frac{p_{i2}}{p_{.2}} \right) - 1$$

y

$$p_{i1} + p_{i2} = p_{i.} \Rightarrow \begin{cases} \dfrac{p_{i1}}{p_{i.}} = 1 - \dfrac{p_{i2}}{p_{i.}} \\ \\ \dfrac{p_{i2}}{p_{i.}} = 1 - \dfrac{p_{i1}}{p_{i.}} \end{cases}$$

entonces,

$$\sum_i \left(\frac{p_{i1}}{p_{i.}}\right)\left(\frac{p_{i1}}{p_{.1}}\right) + \sum_i \left(\frac{p_{i2}}{p_{i.}}\right)\left(\frac{p_{i2}}{p_{.2}}\right) - 1 = \sum_i \frac{p_{i1}}{p_{.1}}\left(1 - \frac{p_{i2}}{p_{i.}}\right) + \sum_i \frac{p_{i2}}{p_{.2}}\left(1 - \frac{p_{i1}}{p_{i.}}\right) - 1$$

$$= \frac{p_{.1}}{p_{.1}} - \sum_i \frac{p_{i1}\,p_{i2}}{p_{i.}\,p_{.1}} + \frac{p_{.2}}{p_{.2}} - \sum_i \frac{p_{i1}\,p_{i2}}{p_{i.}\,p_{.2}} - 1$$

$$= 1 - \sum_i p_{i1}\,p_{i2}\left(\frac{1}{p_{.1}} + \frac{1}{p_{.2}}\right) = 1 - \sum_i \frac{p_{i1}\,p_{i2}}{p_{i.}}\left(\frac{p_{.2} + p_{.1}}{p_{.1}\,p_{.2}}\right) = 1 - \sum_i \frac{p_{i1}\,p_{i2}}{p_{i.}\,p_{.1}\,p_{.2}}$$

Por lo tanto,

$$\Phi^2 = 1 - \sum_i \frac{p_{i1}\,p_{i2}}{p_{i.}\,p_{.1}\,p_{.2}}$$

□

1.7 Varianza Asintótica.

La expresión para la varianza asintótica de $\hat{\Phi}^2$ es (Bishop, 1975):

$$\mathrm{Var}(\hat{\Phi}^2) = \frac{1}{n}\left\{4\sum_i\sum_j\left(\frac{p_{ij}^3}{p_{i.}^2\,p_{j}^2}\right) - 3\sum_i \frac{1}{p_{i.}}\left(\sum_j \frac{p_{ij}^2}{p_{i.}\,p_{j}}\right)^2 - 3\sum_j \frac{1}{p_{j}}\left(\sum_i \frac{p_{ij}^2}{p_{i.}\,p_{j}}\right)^2 + 2\sum_i\sum_j\left[\frac{p_{ij}}{p_{i.}\,p_{j}}\left(\sum_k \frac{p_{kj}^2}{p_{k.}\,p_{j}}\right)\left(\sum_l \frac{p_{il}^2}{p_{i.}\,p_{l}}\right)\right]\right\}$$

Notar que la varianza asintótica de $\hat{\Phi}^2$ es desconocida porque depende de las probabilidades poblacionales. Su estimador se obtiene aplicando el principio de invarianza, es decir, sustituyendo las p_{ij} por sus estimadores de máxima verosimilitud.

2. EL COEFICIENTE DE CONTINGENCIA DE KARL PEARSON.

2.1. Definición.

Con el fin de normar el coeficiente de contingencia en media cuadrática, Pearson propuso la siguiente medida:

$$C = \sqrt{\frac{\Phi^2}{\Phi^2 + 1}}$$

A esta medida de asociación se le conoce como el coeficiente de contingencia de Karl Pearson.

2.2. Propiedades.

2.2.1. El coeficiente de contingencia de Karl Pearson toma valores en el intervalo $[0,1)$.

2.2.2. Cuando las variables son independientes C=o (porque Φ^2 es, en este caso, igual a cero).

2.3. Desventajas.

2.3.1. C solo puede interpretarse cuando las variables son independientes.

2.3.2. En la práctica, C nunca es igual a uno aunque la asociación entre las variables sea perfecta estricta.

2.3.3. El coeficiente de contingencia de Karl Pearson depende de las probabilidades marginales.

2.4. Teorema.

Si las variables A y B tienen ambas I categorías y además están asociadas en forma perfecta estricta entonces, $\lim_{x \to \infty} C = 1$.

Demostración:

Por el teorema 1.4. se tiene que $\Phi^2 = I-1$

$$\Rightarrow \lim_{I\to\infty} C = \lim_{I\to\infty} \sqrt{\frac{\Phi^2}{\Phi^2+1}} = \lim_{I\to\infty} \sqrt{\frac{I-1}{I}} = \lim_{I\to\infty} \sqrt{1-\frac{1}{I}} = 1$$

Entonces, $\lim_{I\to\infty} C = 1$

□

Este teorema indica que solo teóricamente el coefi-- ciente de contingencia de Karl Pearson toma el valor de 1.

2.5 El Estimador y su Desviación Estándar Asintótica.

El estimador de máxima verosimilitud de C es:

$$\hat{C} = \sqrt{\frac{\hat{\Phi}^2}{\hat{\Phi}^2+1}} = \sqrt{\frac{X^2}{X^2+n}}$$

y la desviación estándar asintótica de $\hat{C}$ es - (Bishop, 1975):

$$\sqrt{\text{Var}\,(\hat{C})} = \frac{\sqrt{\text{Var}\,(\hat{\Phi}^2)}}{2\,\Phi\,(1-\Phi^2)^{3/2}}$$

3. EL INDICE DE TSCHUPROV.

3.1. Definición.

Otra medida que se definió para tratar de normar la Φ^2 es el índice de Tschuprov:

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(I-1)\,(J-1)}}}$$

3.2. Propiedades.

3.2.1. El índice de Tschuprov toma valores en el intervalo $[0, 1]$.

3.2.2. $T = 0$ si las dos variables bajo estudio son independientes.

3.2.3. Si la tabla de contingencia es de I X I y la asociación entre las variables es perfecta estricta, entonces $T = 1$ (esta propiedad es una consecuencia del teorema 1.4.).

3.2.4. Si la tabla es de I X J y la asociación entre las variables es perfecta implícita entonces $T < 1$ (esto es una consecuencia del teorema 1.5.).

3.3. Desventajas.

3.3.1. El índice de Tschuprov depende de las probabilidades marginales de la tabla.

3.3.2. T solo puede ser interpretada cuando es igual a cero o uno.

3.4. El Estimador y su Desviación Estándar Asintótica.

Por el principio de invarianza, el estimador máximo verosímil de T se obtiene sustituyendo Φ^2 por $\hat{\Phi}^2$; es decir,

$$\hat{T} = \sqrt{\frac{\hat{\Phi}^2}{\sqrt{(I-1)\cdot(J-1)}}} = \sqrt{\frac{X^2}{n\sqrt{(I-1)\cdot(J-1)}}}$$

La desviación estándar asintótica de $\hat{T}$ es (Bishop, - 1975 y Kendall, 1963):

$$\sqrt{\text{Var}\ (\hat{T})} = \frac{\sqrt{\text{Var}\ (\hat{\Phi}^2)}}{2T\ (I-1)\cdot(J-1)}$$

4. LA V DE CRAMER.

La V de Cramér, a diferencia del índice de Tschuprov, alcanza su máximo valor cuando la asociación entre las variables es perfecta estricta o perfecta implícita.

4.1. Definición.

La V de Cramér se define como:

$$V = \sqrt{\frac{\Phi^2}{m}}$$

en donde $m = \text{mín}\ \{I-1,\ J-1\}$.

4.2. Propiedades.

4.2.1. La V de Cramér es una medida de asociación que - toma valores en el intervalo $[0,1]$.

4.2.2. Cuando A y B son independientes, $V = 0$

4.2.3. Si A y B están asociadas en forma perfecta estricta o perfecta implícita entonces $V = 1$

4.3. Desventajas.

4.3.1. La medida de asociación V depende de las probabilidades marginales de la tabla.

4.3.2. Solo se puede interpretar cuando es igual a cero o uno.

De acuerdo con el teorema siguiente, el índice de Tschuprov es menor o igual a la V de Cramér; de hecho, cuando la tabla es cuadrada T = V y en cualquier otra tabla T < V.

4.4. Teorema.

El índice de Tschuprov siempre es menor o igual a la V de Cramér.

Demostración.

Caso A. La tabla de contingencia es de I X I.

Como I = J entonces, $\sqrt{(I-1)\,(J-1)} = I-1$

y $m = \min\{I-1, J-1\} = I-1$

$$\Rightarrow T = \sqrt{\frac{\Phi^2}{\sqrt{(I-1)\,(J-1)}}} = \sqrt{\frac{\Phi^2}{m}} = V$$

Es decir, T = V

Caso B. La tabla es de I X J ($I \neq J$).

a. Si $I < J$ entonces,

$I-1 < J-1 \Rightarrow (I-1)^2 < (I-1)\,(J-1) \Rightarrow I-1 < \sqrt{(I-1)\,(J-1)}$

además $m = \min\{I-1, J-1\} = I-1$

$\Rightarrow m < \sqrt{(I-1)\,(J-1)}$

b. Si $J < I$ también es cierto que

$m < \sqrt{(I-1)\,(J-1)}$

Entonces, si $I \neq J$, $m < \sqrt{(I-1)\ (J-1)}$ y $T < V$

□

4.5 El Estimador y su Desviación Estándar Asintótica.

El estimador de máxima verosimilitud de V es:

$$\hat{V} = \sqrt{\frac{\hat{\Phi}^2}{m}} = \sqrt{\frac{X^2}{nm}}$$

y su desviación estándar asintótica es (Bishop, 1975 y - Kendall, 1963):

$$\sqrt{\text{Var}\ (\hat{V})} = \frac{\sqrt{\text{Var}\ (\hat{\Phi}^2)}}{2\ V\ \sqrt{m}} \quad \text{en donde } m = \text{mín}\ \{I-1,\ J-1\}.$$

5. EL COEFICIENTE DE CORRELACION.

El coeficiente de correlación es un índice que permite - calcular el grado de asociación entre dos variables nominales y dicotómicas. Este índice fue diseñado tomando como base la definición del coeficiente de correlación de Pearson. A continuación se da la expresión del coeficiente de correlación y se justifica dicha expresión.

5.1. Definición.

El coeficiente de correlación para dos variables nominales y dicotómicas es:

$$\rho = \frac{p_{11}\,p_{22} - p_{12}\,p_{21}}{\sqrt{p_{1\cdot}\,p_{2\cdot}\,p_{\cdot 1}\,p_{\cdot 2}}}$$

Justificación.

De manera general, la forma de una tabla de contingencia con dos variables nominales y dicotómicas es:

Tabla 5.1.

	B_1	B_2	Totales
A_1	p_{11}	p_{12}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	$p_{2\cdot}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	1

Si las categorías A_1 y B_1 se codifican como cero y las categorías A_2 y B_2 como 1 se tendrá que las variables A y B se distribuyen, ambas, como una Bernoulli, con parámetros $p_{2\cdot}$ y $p_{\cdot 2}$ respectivamente. Además,

$$E(A) = p_{2\cdot} \quad ; \quad Var(A) = p_{2\cdot}(1-p_{2\cdot}) = p_{2\cdot}\, p_{1\cdot}$$

$$E(B) = p_{\cdot 2} \quad y \quad Var(B) = p_{\cdot 2}(1-p_{\cdot 2}) = p_{\cdot 2}\, p_{\cdot 1}$$

Por otra parte, la variable AB es igual a uno cuando A y B son iguales a uno y es igual a cero en cualquier otro caso. En consecuencia, $E(AB) = \sum_{c=0}^{1} cP(AB=c) = P(A=1 \text{ y } B=1) = p_{22}$. Por lo tanto,

$$\rho = \frac{E(AB) - E(A)\,E(B)}{\sqrt{Var(A)\;\;Var(B)}} = \frac{p_{22} - p_{2\cdot}p_{\cdot 2}}{\sqrt{p_{1\cdot}p_{2\cdot}p_{\cdot 1}\,p_{\cdot 2}}} \qquad (E\text{-}5.1.)$$

en donde:

$$p_{22} - p_{2\cdot}\,p_{\cdot 2} = p_{22} - (p_{21} + p_{22})\,(p_{12} + p_{22}) = p_{22} - (p_{12}\,p_{21} + p_{21}\,p_{22} + p_{12}\,p_{22} + p_{22}^2)$$

$$= p_{22}\,(1 - p_{21} - p_{12} - p_{22}) - p_{12}\,p_{21} = p_{22}\,p_{11} - p_{12}\,p_{21}$$

$$\Rightarrow \quad \rho = \frac{p_{11}\,p_{22} - p_{12}\,p_{21}}{\sqrt{p_{1\cdot}\,p_{2\cdot}\,p_{\cdot 1}\,p_{\cdot 2}}} \quad .$$

5.2. Interpretación.

El coeficiente de correlación elevado al cuadrado se interpreta de la misma forma que el coeficiente de determinación. Es decir, si A es la variable dependiente y B es la variable independiente ρ^2 indica qué proporción de la variabilidad de A se explica por la variabilidad de B. ρ^2 no tiene una interpretación cuando A y B son simétricas.

5.3 Propiedades.

5.3.1. El coeficiente de correlación toma valores en el intervalo $[-1, 1]$.

5.3.2. ρ es igual a 1 ó -1 si la asociación entre A y B es perfecta estricta. De hecho, $\rho = 1$ si $p_{12} = p_{21} = 0$ y $\rho = -1$ cuando $p_{11} = p_{22} = 0$

5.3.3. Cuando las variables son independientes $\rho = 0$ (ya que en la expresión (E-5.1.) se tendría que $p_{22} - p_{2\cdot}\,p_{\cdot 2} = p_{2\cdot}\,p_{\cdot 2} - p_{2\cdot}\,p_{\cdot 2} = 0$).

5.4. Teorema.

El signo de ρ se modifica al intercambiar los renglones - (o las columnas) de una tabla de 2 X 2. Es decir, si en la tabla 5.1. $\rho = \dfrac{p_{11}\, p_{22} - p_{12}\, p_{21}}{\sqrt{p_{1\cdot} p_{2\cdot} p_{\cdot 1} p_{\cdot 2}}}$

entonces, en la tabla 5.4. $\rho = \dfrac{p_{12}\, p_{21} - p_{11}\, p_{22}}{\sqrt{p_{1\cdot} p_{2\cdot} p_{\cdot 1} p_{\cdot 2}}}$

Tabla 5.4

	B_1	B_2	Totales
A_2	p_{21}	p_{22}	$p_{2\cdot}$
A_1	p_{11}	p_{12}	$p_{1\cdot}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	1

Cuando las variables en estudio son nominales, el orden - de las categorías así como el signo de ρ son irrelevantes.

5.5. El Estimador y su Desviación Estándar Asintótica.

El estimador de máxima verosimilitud de ρ se denota como $\hat{\rho}$ o r y es igual a:

$$\hat{\rho} = \frac{N_{11}\, N_{22} - N_{12}\, N_{21}}{\sqrt{N_{1\cdot} N_{2\cdot} N_{\cdot 1} N_{\cdot 2}}}$$

Su varianza asintótica es (Bishop, 1975; Kendall, 1963; - Reynolds, 1977):

$$Var(r) = \frac{1}{n}\left\{1-\rho^2+\left(\rho+\frac{\rho^3}{2}\right)\frac{(p_{1.}-p_{2.})(p_{.1}-p_{.2})}{p_{1.}p_{2.}p_{.1}p_{.2}} - \frac{3}{4}\rho^2\left[\frac{(p_{1.}-p_{2.})^2}{p_{1.}p_{2.}}+\frac{(p_{.1}-p_{.2})^2}{p_{.1}p_{.2}}\right]\right\}$$

Cuando las variables son independientes (es decir, si $\rho=0$) Var (r) = $\frac{1}{n}$. Además, si en una tabla, todas las probabilidades marginales son iguales a un medio (usualmente las tablas estandarizadas se presentan de esta forma), Var (r) = $\frac{1}{n}(1-\rho^2)$.

Es importante aclarar que el coeficiente de correlación no es una medida de asociación basada en la estadística ji-cuadrada pero se incluyó en este Capítulo porque de acuerdo con el teorema siguiente, si la tabla es de 2 X 2, $\rho^2 = \Phi^2$.

5.6 Teorema.

Cuando la tabla de contingencia es de 2 X 2 el coeficiente de correlación elevado al cuadrado (ρ^2) es igual al coeficiente de contingencia en media cuadrática (Φ^2).

Demostración.

Por definición $$\Phi^2 = \sum_i \sum_j \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

y $$\sum_i \sum_j \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_i \sum_j \frac{\left[p_{ij}-(p_{i1}+p_{i2})(p_{1j}+p_{2j})\right]^2}{p_{i.}p_{.j}}$$

$$= \sum_i \sum_j \frac{1}{p_{i.}p_{.j}}\left(p_{ij}-p_{i1}p_{1j}-p_{i1}p_{2j}-p_{i2}p_{1j}-p_{i2}p_{2j}\right)^2$$

$$= \frac{1}{p_{1\cdot}p_{\cdot 1}}\left(p_{11} - p_{11}^2 - p_{11}p_{21} - p_{12}p_{11} - p_{12}p_{21}\right)^2$$

$$+ \frac{1}{p_{2\cdot}p_{\cdot 1}}\left(p_{21} - p_{21}p_{11} - p_{21}^2 - p_{22}p_{11} - p_{22}p_{21}\right)^2$$

$$+ \frac{1}{p_{1\cdot}p_{\cdot 2}}\left(p_{12} - p_{11}p_{12} - p_{11}p_{22} - p_{12}^2 - p_{12}p_{22}\right)^2$$

$$+ \frac{1}{p_{2\cdot}p_{\cdot 2}}\left(p_{22} - p_{21}p_{12} - p_{21}p_{22} - p_{22}p_{12} - p_{22}^2\right)^2$$

$$= \frac{1}{p_{1\cdot}p_{\cdot 1}}\left[p_{11}(1 - p_{11} - p_{21} - p_{12}) - p_{12}p_{21}\right]^2 + \frac{1}{p_{2\cdot}p_{\cdot 1}}\left[p_{21}(1 - p_{11} - p_{21} - p_{22}) - p_{11}p_{22}\right]^2$$

$$+ \frac{1}{p_{1\cdot}p_{\cdot 2}}\left[p_{12}(1 - p_{11} - p_{12} - p_{22}) - p_{11}p_{22}\right]^2 + \frac{1}{p_{2\cdot}p_{\cdot 2}}\left[p_{22}(1 - p_{21} - p_{12} - p_{22}) - p_{12}p_{21}\right]^2$$

$$= \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1\cdot}p_{\cdot 1}} + \frac{(p_{21}p_{12} - p_{11}p_{22})^2}{p_{2\cdot}p_{\cdot 1}} + \frac{(p_{12}p_{21} - p_{11}p_{22})^2}{p_{1\cdot}p_{\cdot 2}} + \frac{(p_{22}p_{11} - p_{12}p_{21})^2}{p_{2\cdot}p_{\cdot 2}}$$

$$= (p_{11}p_{22} - p_{12}p_{21})^2 S$$

en donde,

$$S = \left(\frac{1}{p_{1\cdot}p_{\cdot 1}} + \frac{1}{p_{2\cdot}p_{\cdot 1}} + \frac{1}{p_{1\cdot}p_{\cdot 2}} + \frac{1}{p_{2\cdot}p_{\cdot 2}}\right) = \frac{1}{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}\left(p_{2\cdot}p_{\cdot 2} + p_{1\cdot}p_{\cdot 2} + p_{2\cdot}p_{\cdot 1} + p_{1\cdot}p_{\cdot 1}\right)$$

$$= \frac{1}{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}\left[p_{1\cdot}(p_{\cdot 2} + p_{\cdot 1}) + p_{2\cdot}(p_{\cdot 1} + p_{\cdot 2})\right] = \frac{p_{1\cdot} + p_{2\cdot}}{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}} = \frac{1}{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}$$

Esto implica que $(p_{11}p_{22} - p_{12}p_{21})^2 S = \dfrac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}} = \rho^2$

Por lo tanto, $\Phi^2 = \rho^2$.

□

En virtud de este teorema, el coeficiente de contingencia en media cuadrática se puede interpretar cuando las variables A y B son dicotómicas y asimétricas.

6. EJEMPLO.

El ejemplo que se dá a continuación se obtuvo del libro de Kendall (1963). Se refiere a un estudio en el que el objetivo era determinar si existía alguna relación entre la inoculación contral el cólera (A) y el cólera (B). Para ello se estudiaron 818 personas: 279 fueron inoculadas contra el cólera (grupo A_1) y las restantes 539 no fueron inoculadas (grupo A_2). Después de cierto tiempo a todas las personas se les realizó un examen médico con el fin de determinar quienes padecían la enfermedad (grupo B_1) y quienes no la padecían (grupo B_2). Los datos obtenidos se muestran en la tabla siguiente:

	B_1	B_2	Totales
A_1	276	3	279
A_2	473	66	539

Las frecuencias marginales, así como el valor de las medidas de asociación basadas en la estadística ji-cuadrada son:

FRECUENCIAS MARGINALES.

NIVEL	RENGLONES	COLUMNAS
1	279.0	749.0
2	539.0	69.0

MEDIDAS BASADAS EN LA JI-CUADRADA.

ESTADISTICA JI-CUADRADA:	29.6957
COEFICIENTE DE CONTINGENCIA EN MEDIA CUADRATICA:	0.0363
COEFICIENTE DE CONTINGENCIA DE KARL PEARSON:	0.1872
INDICE DE TSCHUPROV:	0.1905
LA V DE CRAMER:	0.1905

La estadística ji-cuadrada fue igual a 29.6957. Como este número es mayor que el valor en tablas de cualquier ji-cuadrada con un grado de libertad, se concluye que los datos no aportan evidencias para afirmar que A y B son independientes.

Por otro lado, si A es la variable independiente, el coeficiente de contingencia en media cuadrática (ver el teorema 5.6.) indica que la razón por la que un individuo padece o no de cólera queda explicada en un 3.63% por el hecho de haber sido inoculado contra el cólera o de no haber recibido el inóculo. Esto quiere decir que la asociación entre A y B es muy débil. Lo mismo indican las otras medidas basadas en la estadística ji-cuadrada. Esto, en parte se debe a que todos los índices se ven afectados por la distribución tan heterogénea de las frecuencias marginales, sobretodo en las columnas.

7. BIBLIOGRAFIA.

1. Bishop, Y.M.M., Fienberg, S., Holland P.W.:
Discrete Multivariate Analysis: Theory and Practice.
MIT Press, Cambridge Mass., 1975.

2. Goodman, L.A., Kruskal, W.H.:
Measures of association for cross classifications.
Jorunal of the American Statistical Association,
1954; 49: 732 - 763.

Measures of association for cross classifications.
II. Further discussion and references.
Journal of the American Statistical Association,
1959; 54: 123 - 163

3. Kendall, M.G., Stuart, A.:
The Advanced Theory of Statistics.
C. Griffin and Co. Ltd., London, 2nd. Ed.,
Vol. 2: Inference and Relationship, 1963.

4. Reynolds, H.T.:
The Analysis of Cross-Classifications.
The Free Press, New York, 1977.

III. MEDIDAS BASADAS EN LA RELACION DE VENTAJA

1. La relación de ventaja.
2. El logaritmo natural de la relación de ventaja.
3. La Q de Yule y la Y de Yule.
4. Ejemplo.
5. Una interpretación para la Q de Yule.
6. Dos generalizaciones de la relación de ventaja.
   - 6.1 El conjunto base de cocientes de productos cruzados.
   - 6.2 Una generalización propuesta por Jordan.
7. Una medida propuesta por Altham.
8. Bibliografía.

III. MEDIDAS BASADAS EN LA RELACIÓN DE VENTAJA

La relación de ventaja (o el cociente de productos cruzados) es una medida de asociación para tablas de 2 X 2 que permite comparar el comportamiento de dos poblaciones (por ejemplo, hombres y mujeres o casos y controles) respecto a una variable nominal y dicotómica. Dicha medida tiene una interpretación cuando, para cada población, se emplea un esquema de muestreo multinomial. En cualquier otra situación, el cociente de productos cruzados solo es interpretable si las variables son independientes o si están asociadas en forma perfecta.

Las medidas de asociación consideradas en este Capítulo son:

I. Medidas para tablas de 2 X 2.

1. La relación de ventaja.

2. El logaritmo natural de la relación de ventaja.

3. La Q de Yule y la Y de Yule.

II. Medidas para tablas de I X J ($I > 2$ ó $J > 2$).

1. Dos generalizaciones de la relación de ventaja.

2. Una medida propuesta por Altham.

1. LA RELACION DE VENTAJA.

1.1. Introducción.

En un estudio (Fleiss, 1981) se quería saber cuál era la relación que había entre la edad de las mujeres que

III. MEDIDAS BASADAS EN LA RELACIÓN DE VENTAJA

La relación de ventaja (o el cociente de productos cruzados) es una medida de asociación para tablas de 2 X 2 que permite comparar el comportamiento de dos poblaciones (por ejemplo, hombres y mujeres o casos y controles) respecto a una variable nominal y dicotómica. Dicha medida tiene una interpretación cuando, para cada población, se emplea un esquema de muestreo multinomial. En cualquier otra situación, el cociente de productos cruzados solo es interpretable si las variables son independientes o si están asociadas en forma perfecta.

Las medidas de asociación consideradas en este Capítulo son:

I. Medidas para tablas de 2 X 2.

1. La relación de ventaja.
2. El logaritmo natural de la relación de ventaja.
3. La Q de Yule y la Y de Yule.

II. Medidas para tablas de I X J (I > 2 ó J > 2).

1. Dos generalizaciones de la relación de ventaja.
2. Una medida propuesta por Altham.

1. LA RELACION DE VENTAJA.

1.1. Introducción.

En un estudio (Fleiss, 1981) se quería saber cuál era la relación que había entre la edad de las mujeres que

tuvieron un hijo y el peso de éste al nacer. Para evitar que la asociación dependiera de la raza y del nivel socioeconómico de las madres se decidió estudiar únicamente a las mujeres de raza negra que tuvieran un determinado nivel socioeconómico y que dieran a luz en un determinado hospital. La población de madres estudiadas estuvo formada por 50 mujeres con 20 años o menos y 150 con más de 20 años. Los hijos se clasificaron en dos grupos, los que pesaron 2,500 gramos o menos y los que tuvieron un peso mayor. La tabla de contingencia que se obtuvo es:

	Peso Hijo (grs)		
Edad Madre	$\leq$ 2,500	$>$2,500	Totales
$\leq$ 20 años	10	40	50
$>$ 20 años	15	135	150

Como se puede ver, el número de mujeres de 20 años de edad o más jóvenes que tuvieron un hijo cuyo peso fue de 2,500 gramos o menos fue cuatro veces menor que el número de mujeres del mismo grupo de edad que tuvieron un hijo que pesó más de 2,500 gramos. De aquí que la proporción $\frac{n_{11}}{n_{12}}$ sea igual a $\frac{1}{4}$=0.25. En cambio, para las mujeres de más de 20 años de edad, por cada una que tuvo un hijo que pesó cuando mucho 2,500 gramos hubo nueve que tuvieron un hijo de más de 2,500 gramos, como lo indica la proporción $\frac{n_{21}}{n_{22}} = \frac{15}{135} = \frac{1}{9} = 0.11$.

Nótese que $\frac{n_{11}}{n_{12}} \neq \frac{n_{21}}{n_{22}}$ por lo tanto, las variables están - asociadas. Para determinar el grado de asociación entre la - edad de la madre y el peso del hijo se puede calcular la - siguiente medida:

1.2. Definición.

El cociente de productos cruzados o la relación de - ventaja se define como:

$$\alpha = \frac{\frac{n_{11}}{n_{12}}}{\frac{n_{21}}{n_{22}}} = \frac{\frac{p_{11}}{p_{12}}}{\frac{p_{21}}{p_{22}}} = \frac{p_{11}\, p_{22}}{p_{12}\, p_{21}}$$

Esta medida es el cociente de las proporciones calculadas en el ejemplo anterior y es igual al producto - de las probabilidades en la diagonal principal de la tabla, dividido por el producto de las probabilidades en la otra diagonal. Por esto a α se le conoce como el cociente de productos cruzados.

1.3. Interpretación.

1.3.1. Si la relación de ventaja es igual a uno las variables son independientes.

1.3.2. Cuando $\alpha = 0$ puede suceder que la asociación entre las variables sea perfecta estricta (si $p_{11} = p_{22} = 0$) o perfecta débil (si $p_{11} = 0$ ó $p_{22} = 0$).

1.3.3. Cuando α es indeterminada las variables también pueden estar asociadas en forma perfecta estricta (si $p_{12} = p_{21} = 0$) o perfecta débil (si $p_{12} = 0$ ó -

$p_{21} = 0$).

1.3.4. Cualquier otro valor distinto de las anteriores solo puede interpretarse cuando, en la tabla de contingencia, las frecuencias marginales de alguna de las variables están fijas. Además, la interpretación depende de cuáles sean estas frecuencias marginales.

En el ejemplo anterior, el número de mujeres en cada grupo de edad se conocía desde un principio, es decir, los totales por renglón eran fijos (no eran variables aleatorias) y de acuerdo con las proporciones calculadas $\alpha = \frac{\frac{1}{4}}{\frac{1}{9}} = \frac{9}{4} = 2.25$.

Esto quiere decir que la proporción de mujeres de 20 años o menos que tuvieron un hijo que pesó, cuando mucho, 2,500 gramos fue 2.25 veces mayor que la proporción de madres con más de 20 años cuyo hijo no sobrepasó los 2,500 gramos. Por lo tanto, el peso de los hijos está relacionado con la edad de las madres.

En este ejemplo la relación de ventaja permitió comparar la influencia de la edad de la madre en el peso del hijo. Este es un problema muy distinto al que se hubiera tenido si las frecuencias marginales por columna hubieran sido fijas; en este caso el interés sería determinar la influencia del peso de los hijos en la edad de las madres, lo cual carece de sentido. El cociente de productos cruzados sería igual a:

$$\alpha = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{\frac{2}{3}}{\frac{8}{27}} = \frac{9}{4} = 2.25$$

Independientemente del conjunto de frecuencias marginales que se fije de antemano el valor de α es siempre el mismo, pero la interpretación cambia ya que las proporciones $\frac{n_{11}}{n_{21}}$ y $\frac{n_{12}}{n_{22}}$ se interpretan de manera distinta a $\frac{n_{11}}{n_{12}}$ y $\frac{n_{21}}{n_{22}}$.

1.4. Propiedades.

1.4.1. El cociente de productos cruzados toma valores en el intervalo $[0, \infty)$. (Convencionalmente se acepta que la asociación entre las variables es "negativa" si $\alpha < 1$ y es "positiva" si $\alpha > 1$).

1.4.2. α no depende de las probabilidades (o frecuencias) marginales de la tabla; es decir, α no varía al multiplicar los renglones o las columnas de la tabla por una constante distinta de cero. Esta propiedad permite comparar los índices calculados para tablas de contingencia que tengan distintas probabilidades marginales.

1.5. Desventajas.

1.5.1. El valor de α depende del orden en el que se coloquen las categorías en la tabla de contingencia. De hecho, al intercambiar únicamente los -

renglones o únicamente las columnas la relación de - ventaja es igual al inverso del valor calculado para la tabla original. Si se intercambian los renglones y las columnas el valor de α no se altera.

1.5.2. La escala para medir la asociación "negativa" $[0, 1)$, es distinta a la escala para medir la asociación "positiva", $(1, \infty)$; por lo tanto, debe tenerse cuidado al comparar los índices mayores de uno - con los menores de uno.

1.6. El Estimador y su Varianza Asintótica.

El estimador de máxima verosimilitud de α es:

$$\hat{\alpha} = \frac{N_{11}\ N_{22}}{N_{12}\ N_{21}}$$

Cuando todas las probabilidades poblacionales son - positivas, es decir mayores que cero, la varianza - asintótica de $\hat{\alpha}$ es (Bishop, 1975):

$$\text{Var}\ (\hat{\alpha}) = \frac{\alpha^2}{n}\left(\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right)$$

Por el principio de invarianza, el estimador de máxima verosimilitud de la varianza asintótica de $\hat{\alpha}$ es:

$$\widehat{\text{Var}}\ (\hat{\alpha}) = \hat{\alpha}^2\left(\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}\right) \quad \text{si } N_{ij} \neq 0\ \forall\ i,j$$

2. EL LOGARITMO NATURAL DE LA RELACION DE VENTAJA.

Una transformación del cociente de productos cruzados que permite medir, en la misma escala, la asociación "positiva"

y la asociación "negativa" es la siguiente.

2.1. Definición.

El logaritmo natural de la relación de ventaja se - define como:

$$\alpha' = \ln \alpha = \ln \left(\frac{p_{11}\, p_{22}}{p_{12}\, p_{21}} \right) = \ln p_{11} + \ln p_{22} - \ln p_{12} - \ln p_{21}$$

2.2. Propiedades.

2.2.1. El logaritmo natural del cociente de productos cruzados es una medida de asociación que toma - valores en el intervalo $(-\infty, \infty)$.

2.2.2. Si las variables son independientes α' es - igual a cero (porque en este caso $\alpha = 1$).

2.2.3. α', al igual que α, no depende de las probabilidades marginales de la tabla.

2.3. Interpretación.

2.3.1. Si $\alpha' = 0$ entonces las variables en estudio son independientes.

2.3.2. Si α' es indeterminada la asociación entre las variables puede ser perfecta estricta o perfecta débil.

2.4. Desventajas.

Esta medida de asociación solo se puede interpretar cuando es igual a cero o cuando es indeterminada.

2.5. Estimador.

El estimador máximo verosímil de α' es:

$$\hat{\alpha}' = \ln N_{11} + \ln N_{22} - \ln N_{12} - \ln N_{21}$$

3. LA Q DE YULE Y LA Y DE YULE.

A principios del siglo XX, George Undy Yule propuso el coeficiente siguiente para medir la asociación entre dos variables nominales y dicotómicas:

3.1. Definición.

La Q de Yule se define como:

$$Q=\frac{n_{11}\ n_{22}-n_{12}\ n_{21}}{n_{11}\ n_{22}+n_{12}\ n_{21}}=\frac{p_{11}\ p_{22}-p_{12}\ p_{21}}{p_{11}\ p_{22}+p_{12}\ p_{21}}=\frac{\frac{p_{11}\ p_{22}}{p_{12}\ p_{21}}-\frac{p_{12}\ p_{21}}{p_{12}\ p_{21}}}{\frac{p_{11}\ p_{22}}{p_{12}\ p_{21}}+\frac{p_{12}\ p_{21}}{p_{12}\ p_{21}}}=\frac{\alpha-1}{\alpha+1}$$

3.2. Propiedades.

3.2.1. Q toma valores en el intervalo $[-1, 1]$, (convencionalmente se dice que si $Q < 0$ la asociación es negativa y si $Q > 0$ la asociación es positiva).

3.2.2. Q = 0 si, y solo si, las variables son independientes.

3.2.3. Q es igual a 1 ó -1 cuando la asociación entre las variables es perfecta estricta o perfecta débil.

3.2.4. Q no depende de las probabilidades marginales de la tabla.

Nótese que estas propiedades son similares a las del coeficiente de correlación de Pearson; esta fue la intención de Yule.

Otro índice que propuso Yule para medir la asociación entre dos variables nominales y dicotómicas es el siguiente.

3.3. Definición.

La Y de Yule se define como:

$$Y = \frac{\sqrt{p_{11}\,p_{22}} - \sqrt{p_{12}\,p_{21}}}{\sqrt{p_{11}\,p_{22}} + \sqrt{p_{12}\,p_{21}}} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}$$

Esta medida tiene exactamente las mismas propiedades que Q.

3.4. Interpretación de Q y Y.

3.4.1. Tanto Q como Y son iguales a cero en el caso de independencia.

3.4.2. Si $Q = Y = 1$ y $p_{12} = 0$ ó $p_{21} = 0$ entonces la asociación entre las variables es perfecta - débil.

3.4.3. Si $Q = Y = 1$ y $p_{12} = p_{21} = 0$, la asociación es perfecta estricta.

3.4.4. Si $Q = Y = -1$ y $p_{11} = 0$ ó $p_{22} = 0$ entonces la asociación es perfecta débil.

3.4.5. Si $Q = Y = -1$ y $p_{11} = p_{22} = 0$ la asociación es perfecta estricta.

(Posteriormente se señala cuándo y cómo se puede interpretar cualquier valor distinto de cero, uno o - menos uno).

Los dos teoremas siguientes indican cómo se relaciona la Y de Yule con la Q de Yule y el coeficiente de correlación.

3.5. Teorema.

El valor de la Y de Yule siempre es menor o igual al de la Q de Yule.

(La demostración se basa en las propiedades de la función: raíz cuadrada).

3.6. Teorema.

En una tabla estandarizada de 2 X 2 se tiene que $\rho = Y$.

Demostración.

De acuerdo con el teorema 8.3.3.2.1. del Capítulo I, en una tabla de 2 X 2 estandarizada se tiene que:

$$p_{11} = p_{22} = \frac{\sqrt{\alpha}}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right) \quad ; \quad p_{12} = p_{21} = \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right)$$

$$Y \quad p_{1\cdot} = p_{2\cdot} = p_{\cdot 1} = p_{\cdot 2} = 0.5$$

Esto quiere decir que:

$$\rho = \frac{p_{11}\,p_{22} - p_{12}\,p_{21}}{\sqrt{p_{\cdot 1}\,p_{\cdot 2}\,p_{1\cdot}\,p_{2\cdot}}} = \frac{\frac{\alpha}{4}\left(\frac{1}{\sqrt{\alpha}+1}\right)^2 - \frac{1}{4}\left(\frac{1}{\sqrt{\alpha}+1}\right)^2}{\left(\frac{1}{2}\right)^2}$$

$$= \frac{\alpha - 1}{(\sqrt{\alpha}+1)^2} = \frac{(\sqrt{\alpha}+1)\,(\sqrt{\alpha}-1)}{(\sqrt{\alpha}+1)^2} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1} = Y$$

Por lo tanto, $\rho = Y$.

□

3.7 Corolario.

Si existe una relación de dependencia entre las variables en estudio y la tabla de 2 X 2 es estandarizada entonces la Y de Yule elevada al cuadrado se interpreta de la misma manera que ρ^2.

3.8 Estimadores y Desviaciones Asintóticas.

Los estimadores de máxima verosimilitud de Q y Y son, respectivamente:

$$\hat{Q} = \frac{N_{11} N_{22} - N_{12} N_{21}}{N_{11} N_{22} + N_{12} N_{21}} = \frac{\hat{\alpha} - 1}{\hat{\alpha} + 1}$$

$$\hat{Y} = \frac{\sqrt{N_{11} N_{22}} - \sqrt{N_{12} N_{21}}}{\sqrt{N_{11} N_{22}} + \sqrt{N_{12} N_{21}}} = \frac{\sqrt{\hat{\alpha}} - 1}{\sqrt{\hat{\alpha}} + 1}$$

Los estimadores de la desviación estándar asintótica de $\hat{Q}$ y $\hat{Y}$ son, respectivamente (Bishop, 1975; Reynolds, 1977):

$$\sqrt{\widehat{Var}\,(\hat{Q})} = \frac{1}{2}\left(1-\hat{Q}^2\right)\sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}}$$

$$\sqrt{\widehat{Var}\,(\hat{Y})} = \frac{1}{4}\left(1-\hat{Y}^2\right)\sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}}$$

si $N_{ij} \neq 0$ para toda $i = 1, 2$ y $j = 1, 2$.

4. EJEMPLO.

En el ejemplo del Capítulo anterior (Sección 6) se concluyó que la asociación entre las variables A (inoculación contra el cólera) y B (el cólera) es muy débil porque -

las medidas basadas en la estadística ji-cuadrada dependen de las frecuencias marginales de la tabla. Sin embargo, las medidas de asociación basadas en el cociente de productos cruzados indican que la asociación entre las variables A y B es fuerte.

```
PROGRAMO:   REBECA AGUIRRE HERNANDEZ.
LUGAR:      FACULTAD DE CIENCIAS,UNAM.
TEMA:       MEDIDAS DE ASOCIACION NOMINALES.
FECHA:      JULIO DE 1986.
```

MEDIDAS BASADAS EN LA RELACION DE VENTAJA.

RELACION DE VENTAJA	12.8372
LOGARITMO NATURAL DE LA RELACION DE VENTAJA	2.5523
LA Q DE YULE	0.8555
LA Y DE YULE	0.5636

Si la persona que realizó el estudio determinó previamente el número de personas inoculadas y no inoculadas contra el cólera, entonces la relación de ventaja se interpreta como sigue: la proporción de individuos que fueron inoculados contra el cólera y que no se enfermaron fue 12.83 veces mayor que la proporción de individuos que no fueron inoculados y que padecieron la enfermedad. La Q de Yule es muy cercana a uno y es mayor que la Y de Yule; además, todas las medidas basadas en la relación de ventaja indican que la asociación es positiva.

5. UNA INTERPRETACION PARA LA Q DE YULE.

En un principio la Q de Yule solo podía ser interpretada - en caso de independencia o asociación perfecta. En 1954, Goodman y Kruskal definieron una medida de asociación para dos variables ordinales que tiene una interpretación probabilística y que además es igual a la Q de Yule cuando - la tabla es de 2 X 2.

La medida propuesta por Goodman y Kruskal se denota con la letra griega γ y se conoce como la Gamma de Goodman y - Kruskal. A continuación se explica cuál es el modelo probabilístico usado para definir esta medida cuando $I = J = 2$.

5.1. Modelo Probabilístico.

Se escogen al azar y en forma independiente a dos - individuos de la población en estudio. Después, para el i-ésimo individuo seleccionado (con i=1, 2) se definen las variables a_i y b_i de la manera siguiente:

$$a_i = \begin{cases} 1 & \text{si el i-ésimo individuo pertenece a la clase } A_1 \text{ de la variable A.} \\ 0 & \text{si pertenece a } A_2. \end{cases}$$

$$b_i = \begin{cases} 1 & \text{si el i-ésimo individuo pertenece a la clase } B_1 \text{ de la variable B.} \\ 0 & \text{si pertenece a } B_2. \end{cases}$$

Además se definen las variables a_0 y b_0 como:

$$a_0 = a_1 - a_2$$

$$b_0 = b_1 - b_2$$

Puesto que a_1 y a_2 solo son iguales a cero o uno -- entonces a_0 y b_0 - al igual que $a_0 b_0$ - únicamente - pueden tomar el valor de cero, uno y menos uno.

Caso 1. $a_0 b_0 = 1$

$a_0 b_0$ es igual a uno cuando $a_0 = b_0 = 1$ o bien, - cuando $a_0 = b_0 = -1$.

Subcaso 1.1. Si $a_0 = b_0 = 1$,

entonces $a_1 - a_2 = 1$ y $b_1 - b_2 = 1$

$\Rightarrow a_1 = b_1 = 1$ y $a_2 = b_2 = 0$

Esto quiere decir que el individuo 1 - pertenece a la celda (A_1, B_1) y el individuo 2 a la celda (A_2, B_2).

Subcaso 1.2. Si $a_0 = b_0 = -1$,

entonces $a_1 - a_2 = -1$ y $b_1 - b_2 = -1$

$\Rightarrow a_1 = b_1 = 0$ y $a_2 = b_2 = 1$

Esto significa que el primer individuo seleccionado pertenece a la celda (A_2, B_2) y el segundo a la celda (A_1, B_1).

Por lo tanto, $a_0 b_0 = 1$ si los dos individuos elegidos están clasificados en celdas distintas sobre la diagonal - principal de la tabla de 2 X 2.

Gráficamente, la tabla se expresa como sigue:

	B_1	B_2
A_1		0
A_2	0	

Cuando se tiene una tabla de esta forma se dice que los individuos seleccionados son concordantes. A la probabilidad de que dos individuos sean concordantes se le denotará como π_c.

Caso 2. De manera similar, se puede verificar que $a_o b_o = -1$ si uno de los individuos seleccionados pertenece a la celda (A_1, B_2) y el otro a la celda (A_2, B_1). Gráficamente, la tabla de contingencia se presenta así:

	B_1	B_2
A_1	0	
A_2		0

En este caso se dice que los individuos son discordantes. π_d denotará la probabilidad de que dos individuos sean discordantes.

Caso 3. Cuando $a_o b_o = 0$ los dos individuos están clasificados en la misma categoría de A o B. La probabilidad de que esto suceda se denotará como π_e.

Como los tres casos anteriores agotan todas las posibilidades de clasificación de dos individuos en una tabla de 2 X 2 entonces $\Pi_c + \Pi_d + \Pi_e = 1$; esto implica que $\Pi_c + \Pi_d = 1 - \Pi_e$. Además $\frac{\Pi_c}{1-\Pi_e}$ es la probabilidad de que los dos individuos seleccionados sean concordantes dado que su clasificación en A y B no coincide y $\frac{\Pi_d}{1-\Pi_e}$ es la probabilidad de que los dos individuos sean discordantes dado que su clasificación en A y B no es la misma.

En base a estas dos razones se define la siguiente medida de asociación para dos variables ordinales.

5.2. Definición.

La Gamma de Goodman y Kruskal es igual a:

$$\gamma = \frac{\Pi_c - \Pi_d}{1 - \Pi_e} = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}$$

5.3. Interpretación.

Esta medida indica qué tanto más probable es que los dos individuos, seleccionados al azar, sean concordantes y no discordantes dado que su clasificación en A y B no coincide.

El teorema siguiente señala que cuando I = J = 2, se cumple que $\gamma = Q$.

5.4. Teorema.

La Gamma de Goodman y Kruskal es igual a la Q de Yule cuando la tabla de contingencia es de 2 X 2.

Demostración.

En una tabla como la siguiente,

	B_1	B_2
A_1	p_{11}	p_{12}
A_2	p_{21}	p_{22}

se tiene que: $\Pi_c = P(a_o b_o = 1) = 2p_{11}\, p_{22}$ y

$$\Pi_d = P(a_o b_o = -1) = 2p_{12}\, p_{21}$$

$$\Rightarrow \gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d} = \frac{2(p_{11}\, p_{22} - p_{12}\, p_{21})}{2(p_{11}\, p_{22} + p_{12}\, p_{21})} = Q$$

Entonces $\gamma = Q$. □

Como consecuencia de este teorema, la Q de Yule se puede interpretar de la misma manera que la γ de Goodman y Kruskal.

5.5. Ejemplo.

El ejemplo siguiente trata sobre un estudio que se realizó en Estados Unidos con el fin de averiguar qué factores influyen en el voto de una persona.

Las variables en estudio, así como sus respectivas categorías fueron:

A: Partido por el que votó el individuo.

A_1 : Partido Demócrata.

A_2 : Partido Republicano.

B: Partido con el que simpatiza el individuo.

B_1 : Partido Demócrata.

B_2 : Partido Republicano.

La muestra obtenida estuvo integrada por 2191 personas. - La tabla obtenida, así como las medidas basadas en la relación de ventaja son:

	B_1	B_2
A_1	1103	115
A_2	207	766

```
.....................................................
PROGRAMO:    REBECA AGUIRRE HERNANDEZ.
LUGAR:       FACULTAD DE CIENCIAS,UNAM.
TEMA:        MEDIDAS DE ASOCIACION NOMINALES.
FECHA:       JULIO DE 1986.
.....................................................
```

MEDIDAS BASADAS EN LA RELACION DE VENTAJA.

RELACION DE VENTAJA	35.4925
LOGARITMO NATURAL DE LA RELACION DE VENTAJA	3.5693
LA Q DE YULE	0.9452
LA Y DE YULE	0.7125

La Q de Yule resultó ser igual a 0.9452. Es decir, si - dos individuos seleccionados al azar simpatizan con partidos - políticos opuestos y votan por partidos distintos, la probabilidad de que voten por el partido con el que simpatizan excede en 0.9452 a la probabilidad de que voten por el partido con el que no se identifican.

6. DOS GENERALIZACIONES DEL COCIENTE DE PRODUCTOS CRUZADOS.

6.1. EL CONJUNTO BASE DE COCIENTES DE PRODUCTOS CRUZADOS.

6.1.1. Introducción.

Como se verá, no es sencillo obtener, a partir del cociente de productos cruzados una medida de asociación para tablas de I X J que siempre se pueda interpretar.

Una manera de generalizar el cociente de productos cruzados es dividir la tabla de I X J en varias subtablas de 2 X 2 y calcular, para cada subtabla, la relación de ventaja. Por ejemplo, a partir de la tabla 6.1.

Tabla 6.1.

	B_1	B_2	B_3
A_1	p_{11}	p_{12}	p_{13}
A_2	p_{21}	p_{22}	p_{23}
A_3	p_{31}	p_{32}	p_{33}

se pueden obtener nueve subtablas de 2 X 2. A continuación se anotan cuatro de estas subtablas con sus respectivos cocientes:

	B_1	B_2
A_1	p_{11}	p_{12}
A_2	p_{21}	p_{22}

$$\alpha_1 = \frac{p_{11}\, p_{22}}{p_{12}\, p_{21}}$$

	B_1	B_3
A_1	p_{11}	p_{13}
A_2	p_{21}	p_{23}

$$\alpha_2 = \frac{p_{11}\, p_{23}}{p_{21}\, p_{13}}$$

	B_2	B_3
A_1	p_{12}	p_{13}
A_2	p_{22}	p_{23}

$$\alpha_3 = \frac{p_{12}\, p_{23}}{p_{22}\, p_{13}}$$

	B_1	B_2
A_1	p_{11}	p_{12}
A_3	p_{31}	p_{32}

$$\alpha_4 = \frac{p_{11}\, p_{32}}{p_{12}\, p_{31}}$$

Las probabilidades marginales de estas subtablas son variables aleatorias. Esto significa que los cocientes anteriores solo se pueden interpretar cuando las categorías de las subtablas son independientes o cuando están asociadas en forma perfecta estricta o perfecta débil. Otra desventaja de esta generalización es que entre más grande sea la tabla original mayor será el número de subtablas y de subíndices que habrá que calcular.

Este problema en ocasiones, se puede evitar calculando un conjunto base de cocientes. Este conjunto se obtiene fijando una de las celdas de las tablas de 2 X 2. La celda que se fija generalmente es la última y en ella se colocan a los individuos que pertenecen al renglón I y a la columna J de la tabla original, en la siguiente forma:

		B_J
		p_{-J}
A_I	p_{I-}	p_{IJ}

En el primer renglón y en la primera columna se puede colocar cualquier categoría distinta de A_I y B_J respectivamente. Esto significa que I-1 clases pueden ser colocadas en el primer renglón y J-1 en la columna uno. Por lo tanto, al fijar una de las celdas se obtienen (I-1) · (J-1) subtablas como la siguiente:

	B_j	B_J
A_i	p_{ij}	p_{iJ}
A_I	p_{Ij}	p_{IJ}

con $i = 1, \ldots, I-1$

$j = 1, \ldots, J-1$

Con base en este conjunto de subtablas se define la siguiente medida de asociación.

6.1.2. Definición.

Si en la última celda de las tablas 2 X 2 - se coloca a los individuos que pertenecen a las categorías A_I y B_J, el conjunto base de cocientes de productos cruzados, para una tabla de I X J, es:

$$\alpha_{ij} = \frac{p_{ij}\ p_{IJ}}{p_{Ij}\ p_{iJ}}, \quad \text{con} \quad \begin{matrix} i = 1, \ldots, I-1 \\ j = 1, \ldots, J-1 \end{matrix}$$

6.1.3. Interpretación.

Estos cocientes solo pueden interpretarse - cuando:

6.1.3.1. las categorías (de las tablas de 2 X 2) son independientes.

6.1.3.2. las categorías están asociadas en forma perfecta estricta o perfecta débil.

6.1.3.3. $I = 2$, $J > 2$ y en la tabla original el investigador fijó las frecuencias - marginales por renglón.

6.1.3.4. $I > 2$, $J = 2$ y en la tabla original se fijaron las frecuencias marginales por columna.

6.1.4. Desventajas.

Si la dimensión de la tabla original es muy grande el conjunto base de cocientes de productos cruzados también será grande.

6.1.5. Ventajas.

Al dividir una tabla de I X J en varias subtablas de 2 X 2 para calcular el conjunto base de cocientes de productos cruzados se puede determinar qué categorías son independientes y cuáles están asociadas.

6.1.6. Ejemplo.

En la siguiente tabla de 3 X 3 el conjunto base de cocientes de productos cruzados está formado por cuatro índices.

	B_1	B_2	B_3
A_1	17	4	8
A_2	5	12	0
A_3	10	3	13

A continuación se dá el valor de estos índices así como - su logaritmo natural. (La columna titulada SUBINDICES se refiere a los subíndices de cada cociente).

```
.....................................
PROGRAMO:   REBECA AGUIRRE HERNANDEZ.
LUGAR:      FACULTAD DE CIENCIAS,UNAM.
TEMA:       MEDIDAS DE ASOCIACION NOMINALES.
FECHA:      JULIO DE 1986.
.....................................
```

MEDIDAS BASADAS EN LA RELACION DE VENTAJA.

SUBINDICES	RELACION DE VENTAJA	LOGARITMO NATURAL DE LA RELACION DE VENTA..
1,1	2.7625	1.0161
1,2	2.1667	0.7732
2,1	INDEFINIDO	
2,2	INDEFINIDO	

Puesto que α_{21} es indeterminada, las categorías A_2 y A_3 están asociadas en forma perfecta con B_1 y B_3; lo mismo sucede entre A_2 y A_3 respecto a B_2 y B_3.

6.2. UNA GENERALIZACION PROPUESTA POR JORDAN.

6.2.1. Introducción.

Otra manera de generalizar el cociente de - productos cruzados es "colapsar" o resumir en varias tablas de 2 X 2, la información contenida en una - tabla de I X J. Por ejemplo, si a partir de la tabla 6.1. se definen dos nuevas categorías, una para los individuos que fueron clasificados en B_2 y B_3 y

otra para los clasificados en A_2 y A_3 se obtendrá la tabla siguiente:

	B_1	$B_2 \cup B_3$
A_1	p_{11}	$p_{12} + p_{13}$
$A_2 \cup A_3$	$p_{21} + p_{31}$	$p_{22} + p_{23} + p_{32} + p_{33}$

o bien,

	B_1	$\sum_{j \neq 1} B_j$ (*)
A_1	p_{11}	$p_{1\cdot} - p_{11}$
$\sum_{i \neq 1} A_i$	$p_{\cdot 1} - p_{11}$	$1 - p_{1\cdot} - p_{\cdot 1} + p_{11}$

Pero si en lugar de juntar en una sola categoría a los individuos clasificados en A_2 y A_3 se define una categoría para los que pertenecen a A_1 y A_3 se obtendrá la siguiente tabla.

(*) El símbolo $\sum$ denota en esta tabla la unión de conjuntos ajenos.

	B_1	$B_2 + B_3$
A_2	p_{21}	$p_{2\cdot} - p_{21}$
$A_1 + A_3$	$p_{\cdot 1} - p_{21}$	$1 - p_{2\cdot} - p_{\cdot 1} + p_{21}$

En general, lo que se hace es clasificar a la población respecto a dos variables dicotómicas, A y B. Las categorías de A son A_i y su complemento y las de B son B_j y su complemento. La tabla resultante es:

	B_j	$\sum_{l \neq j} B_l$	Totales
A_i	p_{ij}	$p_{i\cdot} - p_{ij}$	$p_{i\cdot}$
$\sum_{k \neq i} A_k$	$p_{\cdot j} - p_{ij}$	$1 - p_{\cdot j} - p_{i\cdot} + p_{ij}$	$1 - p_{i\cdot}$
Totales	$p_{\cdot j}$	$1 - p_{\cdot j}$	1

El número total de tablas que se pueden obtener es I·J puesto que en el primer renglón se pueden colocar I categorías distintas y en la primera columna J. Basado en estas tablas, Jordan propuso el índice siguiente.

6.2.2. Medida Propuesta.

Para cada tabla "colapsada" se calcula el cociente de productos cruzados y después se promedian.

7. UNA MEDIDA PROPUESTA POR ALTHAM.

Edwards (1963) demostró, que para medir la asociación entre dos variables nominales y dicotómicas que se distribuyen conjuntamente como una multinomial, se debe usar una medida que sea función del cociente $\frac{p_{12}\,p_{21}}{p_{11}\,p_{22}}$. Posteriormente, Patricia Altham (1970), basada en el argumento de -

Edwards, concluyó que en una tabla de I X J, las medidas - de asociación deben ser función del conjunto base de cocientes de productos cruzados.

Un ejemplo de las medidas propuestas por Altham es:

$$\left\{ \sum_i \sum_j \left| \log_e \alpha_{ij} \right|^s \right\}^{\frac{1}{s}}$$

con $s \geq 1$ y $\alpha_{ij} = \dfrac{p_{ij}\, p_{IJ}}{p_{iJ}\, p_{Ij}}$ para $i=1, \ldots, I-1$ y $j=1, \ldots, J-1$

Estas medidas tienen la desventaja de que son difíciles de interpretar; además, no se pueden calcular cuando alguna - p_{ij} es igual a cero. Sin embargo, cuando esto último sucede, se recomienda sumar una constante a cada celda con el fin de suavizar la tabla.

8. BIBLIOGRAFIA.

1. Altham, P.M.E.:
   The measurement of association in a contingency table:
   Three extensions of the cross-ratios and metric methods.
   Journal of the Royal Statistical Society.
   1970; Series B, 32: 395-407.
   The measurement of association of rows and columns for an rxs contingency table.
   Journal of the Royal Statistical Society,
   1970; Series B, 32: 63-73.

2. Bishop, Y.M.M., Fienberg, S., Holland P.W.:
   Discrete Multivariate Analysis: Theory and Practice.
   MIT Press, Cambridge Mass., 1975.

3. Edwards, A.W.F.:
The Measure of Association in a 2 X 2 table.
Journal of the Royal Statistical Society,
1963; Series A, 126: 109-114.

4. Fleiss, J.L.:
Statistical Methods for Rates and Proportions.
John Wiley, New York, 1981.

5. Goodman, L.A., Kruskal, W.H.:
Measures of Association for cross classifications.
Journal of the American Statistical Association,
1954; 49: 732-763.
Measures of association for cross classifications.
II: Further discussion and references.
Journal of the American Statistical Association,
1959; 54: 123-163.

6. Goodman, L.A.:
Simultaneous confidence limits for cross-product ratios
in contingency tables.
Journal of the Royal Statistical Society,
1964; Series B, 26: 86-102.

7. Kendall, M.G., Stuart, A.:
The advanced Theory of Statistics.
C. Griffin and Co. Ltd., London, 2nd. Ed.,
Vol 2: Inference and relationship, 1963.

8. Reynolds, H.T.:
The Analysis of Cross-Classifications.
The Free Press, New York, 1977.

9. Yule, G.U.:
Statistical Papers of G. Udny Yule, borne 1871-died 1951.
Selected by A. Stuart and M.G. Kendall.
C. Griffin, London, 1971; p. 7-170.

IV. LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR

IV. LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR

La mayoría de las medidas de asociación propuestas hasta mediados del siglo XX eran índices que solo podían ser interpretados cuando había independencia o asociación perfecta entre las variables. En 1954, Goodman y Kruskal definieron, con base en distintos modelos probabilísticos, varias medidas de asociación para variables nominales que siempre tienen una interpretación en términos del problema en estudio. Para definir estas medidas, Goodman y Kruskal partieron de la idea de que dos variables nominales están asociadas cuando, al saber cómo fue clasificado un individuo en una de las variables se puede predecir a qué categoría de la otra variable pertenece. Por esta razón, los índices que propusieron miden, en un sentido predictivo, el grado de asociación entre dos o más variables. El modelo probabilístico que usaron para definir dichas medidas es el siguiente: se selecciona al azar a un individuo de la población en estudio y con base en el problema se determina si se predice como está clasificado en A o B cuando:

1. No se sabe cómo fue clasificado en la otra variable.
2. Se sabe a qué categoría de la otra variable pertenece.

Después se calculan $P(1)$ y $P(2)$, las probabilidades de realizar una predicción incorrecta en los casos 1 y 2 respectivamente. Posteriormente dichas probabilidades se comparan mediante

el cociente:

$$\frac{P(1) - P(2)}{P(1)} \qquad \ldots(E - 4.1.)$$

Todas las medidas propuestas por Goodman y Kruskal están basadas en la expresión anterior que se interpreta como la proporción de errores que se eliminan al pasar del caso en el que no se sabe como está clasificado el individuo en alguna de las variables al caso en el que si se sabe. Esta interpretación explica porqué los índices que definieron Goodman y Kruskal se conocen como: "medidas basadas en la lógica de reducción proporcional en el error (RPE)".

Estas medidas tienen las propiedades siguientes:

1. Toman valores en el intervalo $[0, 1]$ porque $P(1)$, la probabilidad de no acertar al realizar las predicciones sin información, es mayor o igual a $P(2)$, la probabilidad de efectuar una predicción errónea al tener información.
2. Si $P(2)=0$ entonces $\frac{P(1)-P(2)}{P(1)} = 1$. Esto significa que, al saber cómo fue clasificado el individuo en alguna de las variables se puede predecir, sin error, su clasificación en la otra variable.
3. Si $P(1) = P(2)$ (es decir, si la probabilidad de no acertar es la misma al tener o no información) entonces $\frac{P(1)-P(2)}{P(1)} = 0$. En otras palabras, el cociente (E - 4.1.) es igual a cero cuando no es útil saber

cómo fue clasificado el individuo en alguna de las variables.

4. Por último, el cociente (E-4.1.) no está definido (o es indeterminado) cuando $P(1)=0$.

Goodman y Kruskal también propusieron dos métodos de predicción. En uno, el objetivo es, en un sentido frecuentista, cometer el menor número de errores. Esto quiere decir que las predicciones se efectúan de tal manera que el número de aciertos sea el máximo al repetir una infinidad de veces el proceso de seleccionar al azar a un individuo y predecir su categoría en alguna de las variables. Las medidas de asociación así definidas se conocen como medidas de predicción óptima. El otro método de predicción consiste en reconstruir, en un sentido frecuentista, la distribución de la población en la tabla de contingencia. Los índices definidos de esta forma reciben el nombre de medidas de predicción proporcional.

Las medidas de predicción óptima se discuten en la sección 1 y las medidas de predicción proporcional en la sección 2.

1. MEDIDAS DE PREDICCION OPTIMA.

En el Capítulo I se mencionó que las variables pueden ser clasificadas como simétricas o asimétricas. Cuando son asimétricas la variable independiente se usa para predecir a que clase de la variable dependiente pertenece el individuo seleccionado. Un ejemplo es el siguiente: una com-

pañía que vende computadoras quiere anunciar sus productos en un solo periódico y desea saber cuál es el periódico más leído por sus posibles compradores. En este caso, se pueden definir las variables

A: necesita una computadora y

B: periódico que lee con mayor frecuencia para tratar de averiguar si es posible predecir a qué clase de B pertenece un individuo cuando se conoce su categoría en A. Por lo tanto, A y B son asimétricas; A es la variable independiente y B es la variable dependiente. Cuando las variables son simétricas tiene sentido tratar de averiguar cómo fue clasificado un individuo tanto en A como en B. Las medidas de predicción óptima se dividen en simétricas o asimétricas dependiendo de la relación que exista entre las variables.

1.1. La Lambda Asimétrica.

1.1.1. Hipótesis.

1.1.1.1. Se tienen dos variables nominales A y B.

1.1.1.2. A tiene I categorías y B tiene J.

1.1.1.3. Las variables son asimétricas y A precede a B en algún sentido.

1.1.2. Modelo Probabilístico.

Se selecciona al azar a un individuo de la población en estudio. El objetivo es predecir a qué

clase de B pertenece cuando:

1. No se sabe cómo fue clasificado en A y

2. Se sabe a qué categoría de A pertenece.

En el primer caso, el número de predicciones correctas, en un sentido frecuentista, se maximiza al decir que el individuo pertenece a la clase B_j con la probabilidad marginal más grande. Esto significa que, en el casol, la probabilidad de acertar al hacer la predicción es $p_{\cdot m} = \max_j \{ p_{\cdot j} \}$ y la probabilidad de cometer un error es:

$$P(1) = 1 - \max_j \{ p_{\cdot j} \}$$

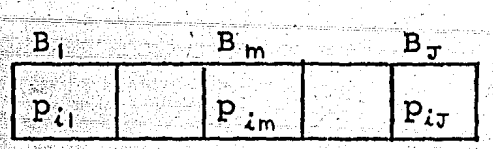
Gráficamente:

	B_1		B_m		B_J
A					
Totales	$p_{\cdot 1}$		$p_{\cdot m}$		$p_{\cdot J}$

$$p_{\cdot m} \geq p_{\cdot j} \quad \forall \; j = 1, \ldots, J.$$

En el segundo caso, se sabe que el individuo seleccionado pertenece a A_i, por lo tanto, el número de predicciones correctas se maximiza si se escoge la categoría B_m que cumple con que $p_{im} = \max_j \{ p_{ij} \}$.

Gráficamente:



$p_{im} \geq p_{ij} \quad \forall \quad j = 1, \ldots, J.$

Como las clases de A son mutuamente excluyentes entonces, al usar la regla 2 la probabilidad de acertar es $\sum_i p_{im}$ y la probabilidad de cometer un error es:

$$P(2) = 1 - \sum_i p_{im}.$$

Si se calcula $\frac{P(1) - P(2)}{P(1)}$ se obtiene la medida siguiente:

1.1.3. Definición.

La lambda asimétrica propuesta por Goodman y Kruskal para predecir categorías de B es:

$$\lambda_j = \frac{P(1)-P(2)}{P(1)} = \frac{1-p_{\cdot m}-1+\sum_i p_{im}}{1-p_{\cdot m}} = \frac{\sum_i p_{im} - p_{\cdot m}}{1-p_{\cdot m}}$$

en donde $p_{\cdot m} = \max_j \{p_{\cdot j}\}$ y $p_{im} = \max_j \{p_{ij}\}$.

1.1.4. Interpretación.

λ_j indica cuanto disminuye la probabilidad relativa de cometer un error al predecir a qué categoría de B pertenece un individuo cuando se pasa del caso en el que no se conoce su clasificación en A al caso en el que si se conoce. Es decir, λ_j indica

cuál es la proporción de errores que se eliminan - cuando se tiene información de A si se compara al - hecho de no tener información de A.

1.1.5. Propiedades.

1.1.5.1. λ_j es indeterminada si, y solo si, la población se concentra en una sola categoría de B.

1.1.5.2. En cualquier otro caso, λ_j es mayor o igual a cero y menor o igual a uno.

1.1.5.3. $\lambda_j = o$ si, y solo si, el saber a - qué clase de A pertenece un individuo no - ayuda a predecir su categoría en B. En otras palabras, $\lambda_j = 0$ si existe alguna j_o tal - que $p_{ij_o} = p_{im}$ para toda i.

1.1.5.4. $\lambda_j = 1$ si, y solo si, la clase a - la que pertenece un individuo en B queda totalmente determinada conociendo su clasificación en A. Es decir, $\lambda_j = 1$ si en cada renglón de la tabla de contingencia hay a lo - más una celda con probabilidad positiva. - (Esto significa que $\lambda_j = 1$ si la asociación entre las variables es perfecta estricta o - perfecta implícita).

1.1.5.5. Si las variables son independientes, $\lambda_j = 0$ pero la inversa no siempre es -

cierta. Es decir, si $\lambda_j = 0$ esto no implica que las variables sean independientes.

1.1.5.6. El valor de λ_j no depende del orden en el que se colocan las categorías en la tabla de contingencia. En otras palabras, λ_j es invariante al intercambiar las columnas y/o los renglones.

Demostración de la propiedad 1.1.5.5.

Parte 1.

Por hipótesis $p_{ij} = p_{i\cdot}p_{\cdot j} \quad \forall \; i,j$

$$\Longrightarrow p_{im} = \max_j \{p_{ij}\} = \max_j \{p_{i\cdot}p_{\cdot j}\} = p_{i\cdot}\max_j \{p_{\cdot j}\}$$

$$= p_{i\cdot}p_{\cdot m}$$

$$\Longrightarrow \lambda_j = \frac{\sum_i p_{im} - p_{\cdot m}}{1-p_{\cdot m}} = \frac{\sum_i p_{i\cdot}p_{\cdot m} - p_{\cdot m}}{1 - p_{\cdot m}} = \frac{p_{\cdot m} - p_{\cdot m}}{1-p_{\cdot m}} = 0$$

$$\therefore \lambda_j = 0$$

Parte 2.

En la tabla siguiente, A y B no son independientes ya que $p_{11} \neq p_{1\cdot}p_{\cdot 1}$

	B_1	B_2	Totales
A_1	0	.75	.75
A_2	.05	.2	.25
Totales	.05	.95	1

Sin embargo, $\lambda_j = \frac{\sum_i p_{im} - p_{\cdot m}}{1 - p_{\cdot m}} = \frac{.95 - .95}{1 - .95} = 0$

□

1.1.6. Desventajas.

La lambda asimétrica depende de las probabilidades marginales de la tabla. Esto quiere decir que los índices calculados para distintas tablas de contingencia solo son comparables cuando las probabilidades marginales son iguales en todas las tablas.

1.1.7. El Estimador y su Varianza Asintótica.

El estimador de máxima verosimilitud de λ_j es:

$$\hat{\lambda}_j = \frac{\sum_i N_{im} - N_{\cdot m}}{n - N_{\cdot m}}$$

en donde $N_{im} = \max_j \{N_{ij}\}$ y $N_{\cdot m} = \max_j \{N_{\cdot j}\}$

La varianza asintótica de $\hat{\lambda}_j$ es (Goodman y Kruskal 1963):

$$\frac{\left(1 - \sum_i p_{im}\right)\left(\sum_i p_{im} + p_{\cdot m} - 2\sum^r p_{im}\right)}{\left(1 - p_{\cdot m}\right)^3}$$

$\sum^r p_{im}$ denota la suma de todas las p_{im} cuya columna coincide con la de $p_{\cdot m}$.

1.1.8 Observaciones.

Es fácil verificar que el índice que se obtiene al predecir renglones (categorías de A) en lugar de columnas (categorías de B) es:

$$\lambda_i = \frac{\sum_j p_{mj} - p_{m\cdot}}{1 - p_{m\cdot}}$$

1.2. La Lambda Condicional.

1.2.1. Introducción.

Para obtener una medida de asociación que no dependa de las probabilidades marginales de la variable A, se pondera cada renglón de la tabla de contingencia por una constante positiva, elegida de tal manera que todas las categorías de A sean equiprobables. Por ejemplo, multiplicando el i-ésimo renglón de la tabla 1.2.1.1. por $\frac{1}{p_{i\cdot} I}$ (para $i = 1, \ldots, I$) se obtiene la tabla 1.2.1.2.; en esta tabla todas las probabilidades marginales de los renglones son iguales a $\frac{1}{I}$.

Tabla 1.2.1.1.

	B_1		B_j		B_J	Totales
A_1	p_{11}		p_{1j}		p_{1J}	$p_{1\cdot}$
A_i	p_{i1}		p_{ij}		p_{iJ}	$p_{i\cdot}$
A_I	p_{I1}		p_{Ij}		p_{IJ}	$p_{I\cdot}$
Totales	$p_{\cdot 1}$		$p_{\cdot j}$		$p_{\cdot J}$	1

Tabla 1.2.1.2.

	B_1		B_j		B_J	Totales
A_1	$\frac{p_{11}}{p_{1\cdot}I}$		$\frac{p_{1j}}{p_{1\cdot}I}$		$\frac{p_{1J}}{p_{1\cdot}I}$	$\frac{1}{I}$
A_i	$\frac{p_{i1}}{p_{i\cdot}I}$		$\frac{p_{ij}}{p_{i\cdot}I}$		$\frac{p_{iJ}}{p_{i\cdot}I}$	$\frac{1}{I}$
A_I	$\frac{p_{I1}}{p_{I\cdot}I}$		$\frac{p_{Ij}}{p_{I\cdot}I}$		$\frac{p_{IJ}}{p_{I\cdot}I}$	$\frac{1}{I}$
Totales	$\frac{1}{I}\sum_i \frac{p_{i1}}{p_{i\cdot}}$		$\frac{1}{I}\sum_i \frac{p_{ij}}{p_{i\cdot}}$		$\frac{1}{I}\sum_i \frac{p_{iJ}}{p_{i\cdot}}$	1

Usando esta última tabla y el modelo probabilístico descrito en 1.1.2. se obtiene que:

I. La probabilidad de predecir incorrectamente a - qué clase de B pertenece un individuo cuando no se sabe cómo fue clasificado en A es:

$$P(1) = 1 - \max_j \left(\frac{1}{I} \sum_i \frac{p_{ij}}{p_{i\cdot}} \right) = 1 - \frac{1}{I} \max_j \left(\sum_i \frac{p_{ij}}{p_{i\cdot}} \right)$$

II. La probabilidad de predecir erróneamente a qué clase de B pertenece un individuo cuando se sabe cómo fue clasificado en A es:

$$P(2) = 1 - \sum_i \max_j \left(\frac{p_{ij}}{p_{i\cdot} I} \right) = 1 - \sum_i \frac{1}{p_{i\cdot} I} \left(\max_j \{ p_{ij} \} \right) = 1 - \sum_i \frac{p_{im}}{p_{i\cdot} I}$$

con $p_{im} = \max_j \{ p_{ij} \}$.

La medida de asociación que resulta al calcular el - cociente (E - 4.1.) es:

1.2.2. Definición.

La lambda condicional para predecir a qué - categoría de B pertenece un individuo se define como:

$$\lambda_j^* = \frac{\frac{1}{I} \sum_i \frac{p_{im}}{p_{i\cdot}} - \frac{1}{I} \max_j \left(\sum_i \frac{p_{ij}}{p_{i\cdot}} \right)}{1 - \frac{1}{I} \max_j \left(\sum_i \frac{p_{ij}}{p_{i\cdot}} \right)}$$

en donde $p_{im} = \max_j \{ p_{ij} \}$.

1.2.3. Comentarios.

1.2.3.1. El índice anterior depende únicamente de las probabilidades marginales de B.

1.2.3.2. De manera similar se puede definir una medida, λ_i^* , cuyo valor solo esté afectado por las probabilidades marginales de A.

1.2.3.3. Estandarizando la tabla de contingencia se puede evitar que las probabilidades marginales de A y B influyan en el valor de las lambdas asimétricas.

1.3. La Lambda Simétrica.

1.3.1. Hipótesis.

1.3.1.1. Se tienen dos variables nominales, A y B.

1.3.1.2. A tiene I categorías y B tiene J.

1.3.1.3. Las variables son simétricas.

1.3.2. Modelo Probabilístico.

Se selecciona al azar a un individuo de la población en estudio y mediante algún procedimiento aleatorio se determina si se va a predecir su categoría en A o su categoría en B. Después se efectúa la predicción usando las dos reglas siguientes:

1. No se sabe como fue clasificado el individuo en la otra variable.

2. Se sabe a qué clase de la otra variable pertenece.

En un sentido frecuentista, el número de predicciones correctas, en el caso uno, se maximiza si al predecir renglones se escoje el que tenga la probabilidad marginal más grande y si al predecir columnas se elije la que tenga la probabilidad marginal mayor. Por lo tanto, la probabilidad de acertar, en el primer caso es:

$$P(\text{acertar con 1}) = P(\text{predecir renglones}) \max_i \{p_{i\cdot}\} + P(\text{predecir columnas}) \max_j \{p_{\cdot j}\}.$$

Si el procedimiento aleatorio es tal que la probabilidad de predecir renglones y columnas es igual a 0.5 y

$$\max_i \{p_{i\cdot}\} = p_{m\cdot} \quad \text{y} \quad \max_j \{p_{\cdot j}\} = p_{\cdot m}$$

entonces, $P(\text{acertar con 1}) = 0.5\,(p_{m\cdot} + p_{\cdot m})$. Esto implica que la probabilidad de cometer un error con la regla 1 es:

$$P(1) = 1 - 0.5\,(p_{m\cdot} + p_{\cdot m}).$$

Al usar la regla 2, la probabilidad de predecir correctamente a qué renglón pertenece el individuo es igual a $\sum_j p_{mj}$ y la probabilidad de predecir correctamente su columna es igual a $\sum_i p_{im}$. Entonces, en el segundo caso, la probabilidad de acertar es:

$$P \text{ (acertar con 2)} = P \text{ (predecir renglones)} \sum_j p_{mj} +$$

$$P \text{ (predecir columnas)} \sum_i p_{im} = 0.5\left(\sum_j p_{mj} + \sum_i p_{im} \right)$$

y la probabilidad de cometer un error es:

$$P\ (2) = 1 - 0.5 \left(\sum_j p_{mj} + \sum_i p_{im} \right) .$$

A continuación se enuncia la medida de asociación que se obtiene al calcular el cociente (E - 4.1.).

1.3.3. Definición.

La lambda simétrica se define como:

$$\lambda = \frac{P\ (1) - P\ (2)}{P\ (1)} = \frac{\sum_j p_{mj} + \sum_i p_{im} - p_{m\cdot} - p_{\cdot m}}{2 - (p_{m\cdot} + p_{\cdot m})}$$

en donde: $p_{m\cdot} = \max_i \{ p_{i\cdot} \}$; $p_{\cdot m} = \max_j \{ p_{\cdot j} \}$;

$p_{mj} = \max_i \{ p_{ij} \}$ y $p_{im} = \max_j \{ p_{ij} \}$.

1.3.4. Interpretación.

La lambda simétrica se interpreta como la proporción de errores que se eliminan al pasar del caso - en el que no se tiene información para hacer la predicción, al caso en el que se cuenta con información.

1.3.5. Propiedades.

1.3.5.1. λ es indeterminada si toda la población se concentra en una sola celda de la tabla de contingencia.

1.3.5.2. Si λ está bien definida (es decir si $p_{m\cdot} + p_{\cdot m} \neq 2$) entonces $0 \leq \lambda \leq 1$.

1.3.5.3. $\lambda = 1$ si, y solo si, en cada renglón y columna hay a lo más una celda cuya probabilidad es mayor que cero. Esto es, - $\lambda = 1$ si, y solo si, la asociación entre - A y B es perfecta estricta.

1.3.5.4. Si A y B son independientes entonces $\lambda = 0$ pero la inversa no siempre es - cierta.

1.3.5.5. λ siempre es menor o igual a una de las lambdas asimétricas y mayor o igual a la otra lambda asimétrica.

1.3.5.6. λ es invariante al permutar los renglones a las columnas de la tabla.

Demostración de la propiedad 1.3.5.5.

Para demostrar la propiedad 1.3.5.5. primero se expresará a λ como función de λ_i y λ_j.

Por definición, $\lambda = \dfrac{\sum_j p_{mj} + \sum_i p_{im} - p_{m\cdot} - p_{\cdot m}}{2 - (p_{m\cdot} + p_{\cdot m})}$

$$\Longrightarrow \lambda = \frac{\left(\sum_i p_{im} - p_{\cdot m}\right) + \left(\sum_j p_{mj} - p_{m\cdot}\right)}{2 - (p_{\cdot m} + p_{m\cdot})} =$$

$$\frac{(1 - p_{\cdot m})\left(\dfrac{\sum_i p_{im} - p_{\cdot m}}{1 - p_{\cdot m}}\right) + (1 - p_{m\cdot})\left(\dfrac{\sum_j p_{mj} - p_{m\cdot}}{1 - p_{m\cdot}}\right)}{(\quad + \quad)}$$

$$\Rightarrow \quad \lambda = \frac{(1 - p_{\cdot m})\lambda_j + (1 - p_{m\cdot})\lambda_i}{2 - (p_{\cdot m} + p_{m\cdot})}$$

Ahora, suponiendo que $\lambda_j \leq \lambda_i$ se demostrará que $\lambda_j \leq \lambda$ y que $\lambda \leq \lambda_i$.

Parte 1: Por demostrar que $\lambda_j \leq \lambda$

Por hipótesis, $\lambda_j \leq \lambda_i \Rightarrow (1-p_{m\cdot})\lambda_j \leq (1-p_{m\cdot})\lambda_i$

$\Leftrightarrow (1-p_{\cdot m})\lambda_j + (1-p_{m\cdot})\lambda_j \leq (1-p_{\cdot m})\lambda_j + (1-p_{m\cdot})\lambda_i$

$\Leftrightarrow [2-(p_{\cdot m} + p_{m\cdot})]\lambda_j \leq (1-p_{\cdot m})\lambda_j + (1-p_{m\cdot})\lambda_i$

$\Leftrightarrow \lambda_j \leq \frac{(1-p_{\cdot m})\lambda_j + (1-p_{m\cdot})\lambda_i}{2 - (p_{\cdot m} + p_{m\cdot})} = \lambda$

$\therefore \lambda_j \leq \lambda$

Parte 2: Por demostrar que $\lambda \leq \lambda_i$

Como $\lambda_j \leq \lambda_i \Rightarrow (1-p_{\cdot m})\lambda_j \leq (1-p_{\cdot m})\lambda_i$

$\Leftrightarrow (1-p_{m\cdot})\lambda_i + (1-p_{\cdot m})\lambda_j \leq (1-p_{m\cdot})\lambda_i + (1-p_{\cdot m})\lambda_i$

$\Leftrightarrow (1-p_{m\cdot})\lambda_i + (1-p_{\cdot m})\lambda_j \leq [2-(p_{m\cdot} + p_{\cdot m})]\lambda_i$

$\Leftrightarrow \lambda = \frac{(1-p_{m\cdot})\lambda_i + (1-p_{\cdot m})\lambda_j}{2 - (p_{m\cdot} + p_{\cdot m})} \leq \lambda_i$

$\therefore \lambda \leq \lambda_i$

Entonces, si $\lambda_j \leq \lambda_i$ se cumple que $\lambda_j \leq \lambda \leq \lambda_i$.

De manera análoga se demuestra que $\lambda_i \leq \lambda \leq \lambda_j$ cuando $\lambda_i \leq \lambda_j$.

□

1.3.6. Desventajas.

La lambda simétrica, al igual que λ_i y λ_j, - depende de las probabilidades marginales de la tabla.

El teorema siguiente indica que, en una tabla de 2 X 2 estandarizada, las dos lambdas asimétricas son iguales a λ y ésta, a su vez, es igual al - valor absoluto de la Y de Yule. Por lo tanto, cuando todas las probabilidades marginales de una tabla de 2 X 2 son iguales a un medio, el valor absoluto - de la Y de Yule se puede interpretar de tres maneras distintas.

1.3.7. Teorema.

En una tabla de 2 X 2 estandarizada se cumple que:

$$\lambda_i = \lambda_j = \lambda = |Y|$$

Demostración.

De acuerdo con el teorema 8.3.3.2.1. del Capítulo I:

$$p_{11} = p_{22} = \frac{\sqrt{\alpha}}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right) \text{ y } p_{12} = p_{21} = \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}+1}\right)$$

$$\Longrightarrow \quad p_{1\cdot} = p_{2\cdot} = p_{\cdot 1} = p_{\cdot 2} = 0.5$$

Caso 1: $\alpha \geq 1$.

Parte 1: Por demostrar que $\lambda_i = |Y|$.

Por definición, $\lambda_i = \dfrac{\sum_j p_{mj} - p_{m\cdot}}{1 - p_{m\cdot}} = \dfrac{p_{m1} + p_{m2} - p_{m\cdot}}{1 - p_{m\cdot}}$

pero, $p_{m\cdot} = \max\{p_{1\cdot}, p_{2\cdot}\} = 0.5 \Longrightarrow 1 - p_{m\cdot} = 0.5$

Como $\alpha \geq 1 \Longrightarrow p_{m1} = \max\{p_{11}, p_{21}\} = p_{11}$

y $p_{m2} = \max\{p_{12}, p_{22}\} = p_{22}$

Además, $p_{m1} = p_{m2} = p_{11}$

Entonces, $\lambda_i = \dfrac{2p_{11} - 0.5}{0.5} = 4p_{11} - 1 = \dfrac{4\sqrt{\alpha}}{2}\left(\dfrac{1}{\sqrt{\alpha}+1}\right) - 1$

$$= \frac{2\sqrt{\alpha} - \sqrt{\alpha} - 1}{\sqrt{\alpha}+1} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha}+1} = Y = |Y|.$$

$\therefore \lambda_i = |Y|$.

Parte 2: Demanera similar se comprueba que $\lambda_j = |Y|$

Parte 3: Por demostrar que $\lambda = |Y|$

Como $\lambda = \dfrac{(1-p_{\cdot m})\lambda_j + (1-p_{m\cdot})\lambda_i}{2 - (p_{\cdot m} + p_{m\cdot})} = \dfrac{0.5\lambda_j + 0.5\lambda_i}{2 - 1}$

$$= 0.5\,(\lambda_i + \lambda_j)$$

$= (0.5) \cdot 2\,|Y|$ por las partes 1 y 2.

$\therefore \lambda = |Y|$

Entonces, si $\alpha \geq 1$ y la tabla está estandarizada se cumple que

$$\lambda_i = \lambda_j = \lambda = |Y|$$

Caso 2: $\alpha < 1$ (La demostración es similar a la anterior).

□

El corolario siguiente es una consecuencia del teorema - que se acaba de enunciar y del teorema 3.6. del Capítulo III.

1.3.8. Corolario.

En una tabla de 2 X 2 estandarizada, se cumplen las siguientes relaciones:

$$\lambda_i = \lambda_j = \lambda = |\rho|$$

1.3.9. El Estimador de Máxima Verosimilitud.

El estimador de máxima verosimilitud de la - lambda simétrica es:

$$\hat{\lambda} = \frac{\sum_j N_{mj} + \sum_i N_{im} - N_{m\cdot} - N_{\cdot m}}{2n - N_{m\cdot} - N_{\cdot m}}$$

en donde,

$$N_{m\cdot} = \max_i \{ N_{i\cdot} \} \quad , \quad N_{\cdot m} = \max_j \{ N_{\cdot j} \}$$

$$N_{mj} = \max_i \{ N_{ij} \} \quad \text{y} \quad N_{im} = \max_j \{ N_{ij} \}$$

1.3.10. Ejemplo.

La siguiente tabla de contingencia se obtuvo del artículo que Goodman y Kruskal publicaron en 1954. En ella se muestra cómo fueron clasificadas - 6,800 personas de acuerdo al color de sus ojos y de su cabello. Las categorías de las variables son:

A: Color de ojos	B: Color del cabello
A_1: Azules	B_1: Claro
A_2: Grises o Verdes	B_2: Castaño
A_3: Castaños	B_3: Negro
	B_4: Rojo

Tabla 1.3.10.

	B_1	B_2	B_3	B_4
A_1	1768	807	189	47
A_2	946	1387	746	53
A_3	115	438	288	16

Si solo se quiere determinar que tan fuerte es la asociación entre A y B las variables se consideran simétricas y entonces se debe calcular el índice λ. Un estudio más interesante sería tratar de averiguar si es correcta la creencia popular de que existe una relación entre el color del cabello y el color de los ojos. Por ejemplo, se podría tratar de saber si es cierto que las personas de cabello claro tienen en su mayoría, ojos azules o que hay pocas personas con cabello negro que tienen ojos grises o verdes. En este caso, las variables se consideran asimétricas y el objetivo sería saber si el color de los ojos de una persona se puede predecir conociendo el color de su cabello.

A continuación se presentan las frecuencias marginales de la tabla 1.3.10. con algunas medidas de predicción óptima.

FRECUENCIAS MARGINALES.

NIVEL	RENGLONES	COLUMNAS
1	2811.0	2829.0
2	3132.0	2432.0
3	857.0	1223.0
4		116.0

MEDIDAS BASADAS EN LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR.

LA LAMBDA DE GOODMAN Y KRUSKAL.

PREDICCION DE LAS COLUMNAS.	0.1924
PREDICCION DE LOS RENGLONES.	0.2241
PREDICCION DE COLUMNAS O RENGLONES.	0.2076

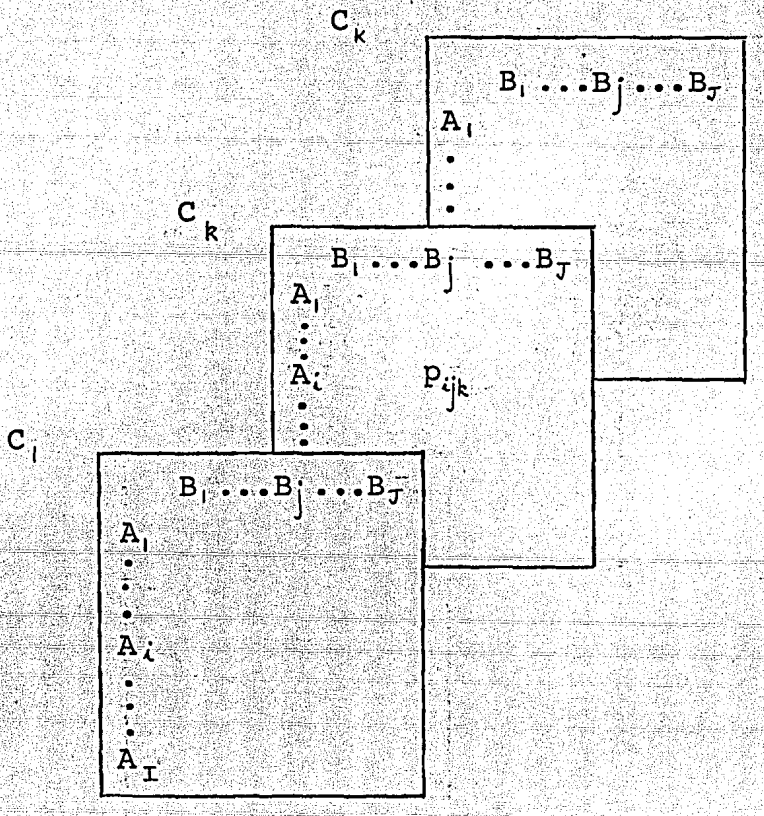
Se puede ver que, en el sentido predictivo, la asociación entre A y B es débil pero debe tenerse en cuenta que todas estas medidas dependen de las frecuencias marginales de la tabla, las cuales son más heterogéneas en las columnas que en los renglones. En parte, esto explica porque λ_j es menor que λ_i .

De acuerdo con λ_i , se reduce en un 22.41 % la probabilidad de predecir erróneamente de qué color son los ojos de una persona cuando se sabe de qué color es su cabello, en contraste a lo que ocurre cuando no se tiene esta información. La lambda simétrica indica que, ya sea que se vaya a predecir el color de los ojos o del cabello, la proporción de errores que se eliminan al pasar del caso en el que no se tiene información, al caso en el que se tiene es igual a 0.2076.

1.4. GENERALIZACIONES PARA TRES VARIABLES NOMINALES.

1.4.1. Introducción.

En esta Sección se definen algunas medidas de predicción óptima para tres variables nominales. Estas variables se denotarán como A, B y C. Se supondrá que cada una de ellas tiene respectivamente las categorías I, J y K. Esto quiere decir que la dimensión de la tabla es I X J X K y gráficamente se puede representar como un conjunto de K tablas con I renglones y J columnas:



Otra manera de presentar una tabla de tres dimensiones es:

C		C_1	...	C_k	...	C_K
B		$B_1 \ldots B_j \ldots B_J$	...	$B_1 \ldots B_j \ldots B_J$	...	$B_1 \ldots B_j \ldots B_J$
A	A_1					
	.					
	.					
	A_i			p_{ijk}		
	.					
	.					
	A_I					

Si un individuo está clasificado en las categorías A_i, B_j y C_k se dirá que pertenece a la celda (A_i, B_j, C_k) y la probabilidad de que un individuo caiga en esta celda se denotará como p_{ijk}. Las probabilidades marginales se denotarán como:

$\sum_j \sum_k p_{ijk} = p_{i..}$ que es la probabilidad de que un individuo pertenezca a la i-ésima categoría de A con $i = 1, \ldots, I$,

$\sum_i \sum_k p_{ijk} = p_{.j.}$ que es la probabilidad de que un individuo esté clasificado en la j-ésima categoría de B (con $j = 1, \ldots, J$) y

$\sum_i \sum_j p_{ijk} = p_{..k}$ que es la probabilidad de que un individuo pertenezca a la k-ésima categoría de C, (con $k = 1, \ldots, k$).

Por último,

$p_{ij\cdot} = \sum_k p_{ijk}$ denotará la probabilidad de que un individuo pertenezca a la i-ésima categoría de A y - j-ésima categoría de B. En otras palabras, $p_{ij\cdot}$ denotará la probabilidad de que un individuo caiga en alguna de las siguientes - celdas: (A_i, B_j, C_1), (A_i, B_j, C_2), ... - (A_i, B_j, C_k), ... (A_i, B_j, C_{k-1}), (A_i, B_j, C_k)

$p_{i\cdot k} = \sum_j p_{ijk}$ denotará la probabilidad de que un individuo esté clasificado en la i-ésima clase de A y k -ésima clase de C.

$p_{\cdot jk} = \sum_i p_{ijk}$ denotará la probabilidad de que un individuo pertenezca a la j-ésima categoría de B y k -ésima categoría de C.

Esta sección se divide en dos partes. En la primera se - definen dos medidas de asociación parcial y en la segunda se - verá una medida de asociación múltiple.

1.4.2. La Asociación Parcial entre A y B.

Cuando se tienen tres variables nominales, la asociación entre dos de ellas se conoce como - asociación parcial. Este tipo de asociación se - puede medir de varias maneras; a continuación se - mencionan dos de ellas.

1.4.2.1. Hipótesis.

1.4.2.1.1. Se tienen tres variables nominales, A, B y C que tiene I, J y K categorías respectivamente.

1.4.2.1.2. A y B son asimétricas y A precede a B en algún sentido.

1.4.2.1.3. Se desea obtener una medida de asociación entre A y B que esté basada en la lógica de RPE.

1.4.2.2. Primer Modelo Probabilístico.

Se selecciona al azar a un individuo de la categoría C_k. Como A precede a B se tratará de predecir a qué clase de B pertenece el individuo cuando:

1. No se sabe cómo fue clasificado en A
2. Se sabe a qué categoría de A pertenece.

Este modelo probabilístico es similar al que se usó para definir a λ_j; la diferencia es que ahora se tiene una variable más y se sabe a qué categoría de esta variable pertenece el individuo. Al calcular las probabilidades de cometer un error con el modelo que se acaba de mencionar, se obtiene la medida:

$$\lambda_j(K) = \frac{\sum_i p_{imk} - p_{\cdot mk}}{1 - p_{\cdot mk}}$$

en donde $p_{imk} = \max_j \{p_{ijk}\}$ y

$$p_{\cdot mk} = \max_j \{p_{\cdot jk}\}$$

Esta medida de asociación se puede calcular para cada clase de C . Después, si $\lambda_j(k)$ se pondera por $p_{\cdot\cdot k}$ $\forall$ k=1,..., k se obtiene la medida que se enuncia a continuación.

1.4.2.3. Primer Indice Propuesto.

Una medida de asociación parcial entre A y B es:

$$\lambda_j(A, B/C) = \sum_k p_{\cdot\cdot k} \lambda_j(k)$$

en donde $$\lambda_j(k) = \frac{\sum_i p_{imk} - p_{\cdot mk}}{1 - p_{\cdot mk}}$$

1.4.2.4. Interpretación.

El índice anterior solo se puede interpretar cuando es igual a uno o cuando A y B son independientes. Si $\lambda_j(A, B/C) = 1$ entonces la asociación entre A y B es perfecta estricta o perfecta implícita; cuando A y B son independientes $\lambda_j(A, B/C) = 0$.

1.4.2.5. Estimador.

El estimador de máxima verosimilitud de $\lambda_j(A, B/C)$ es:

$$\hat{\lambda}_j(A, B/C) = \sum_k \frac{N_{\cdot\cdot k}}{n} \left(\frac{\sum_i N_{imk} - N_{\cdot mk}}{n - N_{\cdot mk}} \right)$$

en donde: $N_{\cdot mk} = \max_j \{N_{\cdot jk}\}$ y $N_{imk} = \max_j \{N_{ijk}\}$

1.4.2.6. Segundo Modelo Probabilístico.

Se escoge al azar a un individuo de la población en estudio y se determina a qué categoría de C pertenece antes de predecir su clase en B considerando que:

1. No se sabe cómo fue clasificado en A
2. Se sabe cómo fue clasificado en A.

Si el individuo seleccionado pertenece a la k-ésima categoría de C, el número de predicciones incorrectas —en el caso uno— se reduce al mínimo si se escoge aquella clase B_m que cumple con que $p_{\cdot mk} = \max_j \{p_{\cdot jk}\}$. Por lo tanto, en el primer caso, la probabilidad de hacer una predicción incorrecta es $P(1) = 1 - \sum_k p_{\cdot mk}$. En el segundo caso, el número de predicciones erróneas se minimiza si la B_m se escoge de tal manera que -

$p_{imk} = \max_j \{p_{ijk}\}$. Esto implica que en el caso 2 la probabilidad de no acertar es: $P(2) = 1 - \sum_i \sum_k p_{imk}$. Al calcular el cociente (E-4.1.) se obtiene la medida que se define en 1.4.2.7.

1.4.2.7. Segundo Indice Propuesto.

Si A, B y C son tres variables nominales, la asociación parcial entre A y B se puede medir de la siguiente forma:

$$\lambda'_j(A, B/C) = \frac{P(1) - P(2)}{P(1)} = \frac{\sum_i \sum_k p_{imk} - \sum_k p_{\cdot mk}}{1 - \sum_k p_{\cdot mk}}$$

en donde $p_{imk} = \max_j \{p_{ijk}\}$ y $p_{\cdot mk} = \max_j \{p_{\cdot jk}\}$.

1.4.2.8. Interpretación.

El índice anterior se interpreta como la proporción de errores que se eliminan al predecir a qué categoría de B pertenece un individuo cuando se pasa del caso en el que se conoce su categoría en C, al caso en el que se sabe cómo fue clasificado en A y C. Como se puede observar - $\lambda'_j(A, B/C)$, a diferencia de $\lambda_j(A, B/C)$, tiene siempre una interpretación.

1.4.2.9. Estimador y Varianza Asintótica.

El estimador de máxima verosimilitud de $\lambda_j'(A, B/C)$ es:

$$\hat{\lambda}_j'(A, B/C) = \frac{\sum\sum_{ik} N_{imk} - \sum_k N_{\cdot mk}}{n - \sum_k N_{\cdot mk}}$$

con $N_{imk} = \max_j \{N_{ijk}\}$ y $N_{\cdot mk} = \max_j \{N_{\cdot jk}\}$

La varianza asintótica de $\hat{\lambda}_j'(A, B/C)$ es (Goodman y Kruskal, 1963):

$$\frac{(1 - \sum\sum_{ik} p_{imk})\ (\sum\sum_{ik} p_{imk} + \sum_k p_{\cdot mk} - 2\sum\sum^{r} p_{imk})}{(1 - \sum_k p_{\cdot mk})^3};$$

$\sum\sum^{r} p_{imk}$ es la suma de todas las p_{imk} que están en la misma columna en la que $p_{\cdot mk}$ alcanza su máximo.

1.4.2.10. Ejemplo.

A finales del siglo XIX, Francis Galton realizó una serie de estudios sobre la herencia. En uno de éstos clasificó a los hijos de 78 familias de acuerdo a las variables:

A: El niño (a) tiene ojos claros (Si, No).

B: A lguno de sus padres tiene ojos claros (Si, No).

C: Alguno de sus abuelos tiene ojos claros (Si, No).

La tabla siguiente muestra los datos que obtuvo (Steel, 1980).

C		Si		No	
B		Si	No	Si	No
A	Si	1928	552	596	508
	No	303	395	225	501

Para determinar cuál es la asociación parcial entre A y B se calcularán $\hat{\lambda}_j(A, B/C)$ y $\hat{\lambda}_j'(A, B/C)$. Como:

$I = J = K = 2$, $n = 5008$, $N_{..1} = 378$, $N_{..2} = 1830$

$N_{1m1} = \max\{N_{111}, N_{121}\} = 1928$, $N_{1m2} = \max\{N_{112}, N_{122}\} = 596$

$N_{2m1} = \max\{N_{211}, N_{221}\} = 395$, $N_{2m2} = \max\{N_{212}, N_{222}\} = 501$

$N_{.m1} = \max\{N_{.11}, N_{.21}\} = 2231$, $N_{.m2} = \max\{N_{.12}, N_{.22}\} = 1009$

entonces, $\hat{\lambda}_j(A, B/C) = 0.635\ (0.033) - 0.365\ (0.022) = 0.029$

y $\hat{\lambda}_j'(A, B/C) = \dfrac{N_{1m1} + N_{1m2} + N_{2m1} + N_{2m2} - N_{.m1} - N_{.m2}}{n - N_{.m1} - N_{.m2}} = \dfrac{180}{1768} = 0.102$

En consecuencia, la asociación parcial entre A y B es débil pero debe recordarse que las medidas calculadas dependen

de las frecuencias marginales de las variables. El índice - $\lambda_j'(A, B/C)$ señala que se reduce en 10.2% la probabilidad de predecir erróneamente si alguno de los padres del niño seleccionado tiene ojos claros cuando se pasa del caso en el que - solo se tiene información de C al caso en el que se tiene información de A y C.

1.4.3. Una Medida de Asociación Múltiple.

1.4.3.1. Introducción.

Cuando se trabaja con variables nominales, la asociación múltiple se refiere a la asociación entre una variable con el resto. Por ejemplo cuando se tienen - tres variables nominales, la asociación - de B con A y C se denomina asociación múltiple. Para medirla Goodman y Kruskal definieron la variable (AC). Las clases de - esta variable son todas las parejas ordenadas que se pueden formar con las categorías de A y C, es decir:

$$
\begin{array}{llll}
(A_1C_1) & (A_1C_2) & \ldots (A_1C_k) & \ldots (A_1C_K) \\
(A_2C_1) & (A_2C_2) & \ldots (A_2C_k) & \ldots (A_2C_K) \\
\vdots & & & \\
(A_iC_1) & (A_iC_2) & \ldots (A_iC_k) & \ldots (A_iC_K) \\
\vdots & & & \\
(A_IC_1) & (A_IC_2) & \ldots (A_IC_k) & \ldots (A_IC_K)
\end{array} \quad (*)
$$

En seguida se describe el modelo probabilístico que usaron Goodman y Kruskal para definir la medida de asociación - múltiple entre B y (AC).

1.4.3.2. Modelo Probabilístico.

Aleatoriamente se selecciona a un individuo de la población en estudio. El objetivo es predecir a qué categoría de B pertenece cuando:

1. No se sabe cómo está clasificado en (AC).
2. Se sabe a qué categoría de (AC) pertenece.

En el caso uno, y en un sentido frecuentista, el número de predicciones incorrectas se minimiza al escoger la categoría B_m que satisface con que $p_{\cdot m \cdot} = \max_j \{p_{\cdot j \cdot}\}$. Por esto, la probabilidad de cometer un error con la regla 1 es $P(1) = 1 - p_{\cdot m \cdot}$. En el segundo caso, si el individuo seleccionado pertenece a la celda $(A_i C_k)$, el número de predicciones correctas se maximiza al escoger aquella clase de B que -

(*) La probabilidad de que un individuo pertenezca a $(A_i C_k)$ se denotará como $p_{i \cdot k}$ y la probabilidad de que un individuo esté clasificado a la vez en B_j y $(A_i C_k)$ se denotará como p_{ijk}.

cumple con que $p_{imk} = \max_j \{p_{ijk}\}$. De - aquí que la probabilidad de cometer un - error con la regla 2 sea $P(2) = 1 - \sum_i \sum_k p_{imk}$. La medida propuesta por Goodman y Kruskal - es:

1.4.3.3. Definición.

Un índice basado en la lógica de reducción proporcional en el error que mide la asociación múltiple entre B y (AC) - es:

$$\lambda_j(B|A\ C) = \frac{\sum_i \sum_k p_{imk} - p_{\cdot m \cdot}}{1 - p_{\cdot m \cdot}}$$

en donde:

$p_{imk} = \max_j \{p_{ijk}\}$ y $p_{\cdot m \cdot} = \max_j \{p_{\cdot j \cdot}\}$.

1.4.3.4. Interpretación.

El coeficiente anterior se interpreta como la proporción de errores que se eliminan al predecir a qué categoría - de B pertenece un individuo cuando se pasa del caso en el que no se sabe a qué clase de (AC) pertenece, al caso en el que si se sabe.

A continuación se enuncia una relación que existe entre la medida de asociación par-

cial $\lambda'_j(A, B/C)$ y el índice de asociación múltiple que se definió. Esta relación es similar a la que existe entre el coeficiente de correlación múltiple. $\rho_{j \cdot ik}$ -que mide la asociación entre B y (AC)- y el coeficiente de correlación parcial $\rho_{ji \cdot k}$ -que mide la asociación entre B y A cuando C es conocida-.

1.4.3.5. Teorema.

Sean A, B y C tres variables nominales. Si

$$\lambda'_j(A, B/C) = \frac{\sum_i \sum_k p_{imk} - \sum_k p_{\cdot mk}}{1 - \sum_k p_{\cdot mk}}$$

mide la asociación parcial entre A y B dado que C es conocida.

$$\lambda''_j(B/A\ \ C) = \frac{\sum_i \sum_k p_{imk} - p_{\cdot m\cdot}}{1 - p_{\cdot m\cdot}}$$

es un índice de asociación múltiple entre B y (AC).

$$\lambda_j(B/C) = \frac{\sum_k p_{\cdot mk} - p_{\cdot m\cdot}}{1 - p_{\cdot m\cdot}}$$

mide la asociación entre B y C suponiendo que C precede a B en algún sentido,

entonces:

$$\left[1 - \lambda_j(B1\ C)\right]\left[1 - \lambda_j'(A,\ B1\ C)\right]$$

$$= 1 - \lambda_j''(B1\ A,\ C)$$

Demostración.

Como

$$\text{I.}\quad 1 - \lambda_j''(B/\ A,\ C) = \frac{1 - p_{\cdot m \cdot} - \sum_i \sum_k p_{imk} + p_{\cdot m \cdot}}{1 - p_{\cdot m \cdot}} = \frac{1 - \sum_i \sum_k p_{imk}}{1 - p_{\cdot m \cdot}}$$

$$\text{II.}\quad 1 - \lambda_j(B/\ C) = \frac{1 - p_{\cdot m \cdot} - \sum_k p_{\cdot mk} + p_{\cdot m \cdot}}{1 - p_{\cdot m \cdot}} = \frac{1 - \sum_k p_{\cdot mk}}{1 - p_{\cdot m \cdot}}$$

$$\text{III.}\quad 1 - \lambda_j'(A,\ B/\ C) = \frac{1 - \sum_k p_{\cdot mk} - \sum_i \sum_k p_{imk} + \sum_k p_{\cdot mk}}{1 - \sum_k p_{\cdot mk}} = \frac{1 - \sum_i \sum_k p_{imk}}{1 - \sum_k p_{\cdot mk}}$$

Entonces:

$$\left[1 - \lambda_j(B/C)\right]\left[1 - \lambda_j'(A,\ B/C)\right] = \frac{1 - \sum_i \sum_k p_{imk}}{1 - p_{\cdot m \cdot}} = 1 - \lambda_j''(B/A,\ C)$$

Por lo tanto, se concluye que:

$$\left[1 - \lambda_j(B/C)\right]\left[1 - \lambda_j'(A,\ B/C)\right] = 1 - \lambda_j''(B/A\ C)$$

□

2. LAS MEDIDAS DE PREDICCION PROPORCIONAL.

En la sección anterior, cuando no se tenía información para hacer la predicción, el individuo seleccionado siempre era asignado a la clase modal de A (o B). Nunca se decía que un individuo pertenecía a una categoría cuya probabilidad marginal fuera menor que $p_{m.}$ (o $p_{.m}$). Un método de predicción distinto al anterior consiste en decir, unas veces, que el individuo pertenece a A_1 (o B_1), otras, que pertenece a A_2 (o B_2), etc. de tal manera que al repetir una infinidad de veces el proceso de: "seleccionar al azar y de manera independiente a un individuo y predecir su categoría en A (o B)" se obtenga que, la proporción de individuos que fueron asignados a A_i (o B_j) sea aproximadamente igual a $p_{i.}$ (o $p_{.j}$) para $i=1,\ldots,I$ y $j = 1,\ldots,J$. Es decir, cuando no se tiene información para hacer la predicción, lo que se busca, en un sentido frecuentista, es obtener la distribución marginal de A o B. Cuando se sabe cómo fue clasificado el individuo en una de las variables, el objetivo es, en un sentido frecuentista, reconstruir la distribución conjunta de A y B. Este método de predicción no garantiza que, a la larga, el número de asignaciones incorrectas sea mínimo. Las medidas de asociación basadas en este método de predicción se conocen como medidas de predicción proporcional y se dividen en simétricas y asimétricas, dependiendo de las variables.

2.1. La Tau Asimétrica.

2.1.1. Hipótesis.

2.1.1.1. Se tienen dos variables nominales que se denotarán como A y B.

2.1.1.2. A tiene I categorías y B tiene J.

2.1.1.3. Las variables son asimétricas y A precede a B en algún sentido.

2.1.2. Modelo Probabilístico.

Se selecciona al azar un individuo de la población en estudio. Como A precede a B, el objetivo es predecir cómo está clasificado el individuo en B cuando se usan las dos reglas siguientes:

1. No se sabe cuál es su categoría en A y la probabilidad de predecir que está en B_j es $p_{\cdot j}$ para $j=1,\ldots,J$.
2. Se sabe que el individuo está clasificado en A_i por lo que la probabilidad de predecir que pertenece a B_j es $\frac{p_{ij}}{p_{i\cdot}}$ para $i=1,\ldots,I$ y $j=1,\ldots,J$.

Considerando que las categorías de B son mutuamente excluyentes, la probabilidad de acertar con la regla 1 es:

$$\sum_j p(\text{el individuo pertenezca a } B_j)\ P(\text{predecir que está en } B_j) = \sum_j p_{\cdot j}^2$$

y la probabilidad de cometer un error es $P(1)=1-\sum_j p_{\cdot j}^2$.

Cuando se sabe que el individuo pertenece a A_i (regla 2), la probabilidad de efectuar una predicción correcta es:

$$\sum_i \sum_j P \text{ (el individuo esté en } A_i \text{ y } B_j) \cdot P \text{ (predecir } B_j \text{ dado } A_i)$$

$$= \sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}}$$

por lo tanto, $P(2)= 1 - \sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}}$ es la probabilidad de equivocarse. El índice que resulta al calcular $\frac{P(1) - P(2)}{P(1)}$ se conoce como la "tau asimétrica".

Cabe señalar que la regla 1 reconstruye la distribución marginal de B porque las predicciones se realizan de manera proporcional al número de individuos en cada B_j y, con la regla 2, se obtiene a la larga la distribución conjunta de A y B porque las predicciones se hacen en forma proporcional al número de individuos en cada celda. Por lo anterior, los índices definidos con base al modelo probabilístico descrito en 2.1.2. reciben el nombre de medidas de predicción proporcional.

2.1.3. Definición.

La tau asimétrica que propusieron Goodman y Kruskal para predecir categorías de B es:

$$\tau_j = \frac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2}{1 - \sum_j p_{\cdot j}^2}$$

2.1.4. Primera Interpretación.

El valor de τ_j indica cuál es la proporción de errores que se eliminan al pasar del caso en el que B_j se predice con probabilidad $p_{\cdot j}$, al caso en el que se predice con probabilidad $\frac{p_{ij}}{p_{i\cdot}}$.

Otra expresión para τ_j en términos de la diferencia entre p_{ij} y $p_{i\cdot}p_{\cdot j}$ es la siguiente.

2.1.5. Teorema.

La medida de asociación, $\tau_j = \dfrac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2}{1 - \sum_j p_{\cdot j}^2}$

se puede expresar como:

$$\frac{\sum_i \sum_j \frac{1}{p_{i\cdot}} (p_{ij} - p_{i\cdot}\, p_{\cdot j})^2}{1 - \sum_j p_{\cdot j}^2}$$

Demostración.

Para demostrar que $\tau_j = \dfrac{\sum_i \sum_j \frac{1}{p_{i\cdot}} (p_{ij} - p_{i\cdot} p_{\cdot j})^2}{1 - \sum_j p_{\cdot j}^2}$

basta demostrar que $\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2 = \sum_i \sum_j \frac{1}{p_{i\cdot}} (p_{ij} - p_{i\cdot} p_{\cdot j})^2$

Como $\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2 = \sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - 2\sum_j p_{\cdot j}^2 + \sum_j p_{\cdot j}^2$

$$= \sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - 2\left(\sum_j p_{\cdot j} \sum_i p_{ij}\right) + \left(\sum_i p_{i\cdot}\right)\left(\sum_j p_{\cdot j}^2\right)$$

$$= \sum_i \sum_j \frac{1}{p_{i\cdot}} \left(p_{ij}^2 - 2p_{ij}p_{i\cdot}p_{\cdot j} + p_{i\cdot}^2 p_{\cdot j}^2\right)$$

$$= \sum_i \sum_j \frac{1}{p_{i\cdot}} \left(p_{ij} - p_{i\cdot}p_{\cdot j}\right)^2$$

Entonces, $\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2 = \sum_i \sum_j \frac{1}{p_{i\cdot}} \left(p_{ij} - p_{i\cdot}p_{\cdot j}\right)^2$

$$\Rightarrow \tau_j = \frac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i\cdot}} - \sum_j p_{\cdot j}^2}{1 - \sum_j p_{\cdot j}^2} = \frac{\sum_i \sum_j \frac{1}{p_{i\cdot}} \left(p_{ij} - p_{i\cdot}p_{\cdot j}\right)^2}{1 - \sum_j p_{\cdot j}^2}$$

□

Como consecuencia del teorema 2.1.5.

$$\tau_j = \frac{\frac{1}{2} \sum_i \sum_j \frac{1}{p_{i\cdot}} \left(p_{ij} - p_{i\cdot}p_{\cdot j}\right)^2}{\frac{1}{2} - \frac{1}{2} \sum_j p_{\cdot j}^2}$$

En esta expresión, el numerador es una suma de cuadrados que puede interpretarse como la variación conjunta de A y B; el denominador indica cuál es la variabilidad de A. Por esto, τ_j también se puede interpretar como sigue:

2.1.6. Segunda Interpretación.

τ_j indica qué proporción de la variabilidad de B queda explicada por la variabilidad en A.

Esta última interpretación es análoga a la

del coeficiente de determinación.

2.1.7. Propiedades.

2.1.7.1. La tau asimétrica toma valores en el intervalo $[0, 1]$.

2.1.7.2. $\tau_j = 0$ si, y solo si, las variables son independientes.

2.1.7.3. $\tau_j = 1$ si, al saber a qué categoría de A pertenece un individuo se puede predeci r, sin error, su clase en B.

2.1.7.4. τ_j es indeterminada cuando la probabilidad marginal de alguna B_j es igual a uno.

2.1.8. El Estimador y su Varianza Asintótica.

El estimador de máxima verosimilitud de τ_j es:

$$\frac{\sum_i \frac{1}{N_{i.}} \sum_j N_{ij}^2 - \frac{1}{n} \sum_j N_{.j}^2}{n - \frac{1}{n} \sum_j N_{.j}^2} = \frac{\frac{1}{2} \sum_i \sum_j \frac{1}{N_{i.}} \left(N_{ij} - \frac{N_{i.} N_{.j}}{n} \right)^2}{\frac{n}{2} - \frac{1}{2n} \sum_j N_{.j}^2}$$

La varianza asintótica de $\hat{\tau}_j$ es (Goodman y Kruskal, 1972):

$$\frac{1}{\delta^4} \sum_i \sum_j p_{ij} \left(\phi_{ij} - \bar{\phi} \right)^2$$

en donde:

$$\phi_{ij} = -2p_{\cdot j}\left(1-\sum_i\sum_j \frac{p_{ij}^2}{p_{i\cdot}}\right)+2\frac{p_{ij}}{p_{i\cdot}}\left(1-\sum_j p_{\cdot j}^2\right)-\left(1-\sum_j p_{\cdot j}^2\right)\sum_{j=1}^{J}\left(\frac{p_{ij}}{p_{i\cdot}}\right)^2$$

$$\phi = \left(1-\sum_j p_{\cdot j}^2\right)\sum_i\sum_j \frac{p_{ij}^2}{p_{i\cdot}} - 2\sum_j p_{\cdot j}^2$$

2.1.9. Comentarios.

2.1.9.1. Si en el modelo probabilístico descrito 2.1.2. se intercambian los papeles de A y B se obtiene la medida:

$$\tau_i = \frac{\sum_j \frac{1}{p_{\cdot j}} \sum_i p_{ij}^2 - \sum_i p_{i\cdot}^2}{1 - \sum_i p_{i\cdot}^2}$$

Las propiedades y la interpretación de este índice son similares a las de τ_j .

2.1.9.2. A partir del estimador $\hat{\tau}_i$ se define la estadística $U^2=(n-1)\cdot(I-1)\,\hat{\tau}_i$ que sirve para probar si dos variables nominales son independientes. Cuando $p_{ij}=p_{i\cdot}.p_{\cdot j}$ $\forall$ i,j, la estadística U^2 se distribuye asintóticamente como una ji-cuadrada con $k = (I-1)\cdot(J-1)$ grados de libertad.

2.2. La Tau Simétrica.

2.2.1. Hipótesis.

2.2.1.1. Se tienen dos variables nominales: A y B.

2.2.1.2. A tiene I categorías y B tiene J.

2.2.1.3. A y B son simétricas.

2.2.2. Modelo Probabilístico.

Se selecciona al azar a un individuo de la población en estudio y aleatoriamente se decide si se predicen renglones (categorías de A) o columnas (categorías de B). Las predicciones se realizan mediante las dos reglas siguientes:

1. Si se predicen renglones, la i-ésima categoría de A se vaticina con probabilidad $p_{i.}$ y al predecir columnas, Bj se vaticina con probabilidad de p.j.
2. Si se sabe que el individuo fue clasificado en A_i, B_j se predice con probabilidad $\frac{p_{ij}}{p_{i.}}$ y cuando se sabe que pertenece a B_j, la clase A_i se predice con probabilidad $\frac{p_{ij}}{p_{.j}}$.

En el caso 1, la probabilidad de hacer una asignación correcta es:

P (predecir A) P (acertar al predecir A) + P (predecir B) P (acertar al predecir B), que es igual a $\frac{1}{2}\left(\sum_i p_{i.}^2 + \sum_j p_{.j}^2\right)$ si la probabilidad de predecir renglones y columnas es igual a un medio. Por lo tanto, la probabilidad de equivocarse al usar l es:

$$P(1) = 1 - \frac{1}{2}\left(\sum_i p_{i\cdot}^2 - \sum_j p_{\cdot j}^2\right)$$

Análogamente, la probabilidad de acertar - con la regla 2 es:

$$\frac{1}{2}\sum_i\sum_j\left(\frac{p_{ij}^2}{p_{\cdot j}} + \frac{p_{ij}^2}{p_{i\cdot}}\right) = \frac{1}{2}\sum_i\sum_j p_{ij}^2\left(\frac{p_{i\cdot} + p_{\cdot j}}{p_{i\cdot}p_{\cdot j}}\right)$$

y la probabilidad de cometer un error es

$$P(2) = 1 - \frac{1}{2}\sum_i\sum_j p_{ij}^2\left(\frac{p_{i\cdot} + p_{\cdot j}}{p_{i\cdot}p_{\cdot j}}\right)$$

La "tau simétrica" se obtiene al calcular - el cociente (E-4.1.).

2.2.3. Definición.

La tau simétrica propuesta por Goodman y - Kruskal se define como

$$\tau = \frac{1\sum_i\sum_j \frac{p_{ij}^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}p_{\cdot j}} - \frac{1}{2}\left(\sum_i p_{i\cdot}^2 + \sum_j p_{\cdot j}^2\right)}{1 - \frac{1}{2}\left(\sum_i p_{i\cdot}^2 + \sum_j p_{\cdot j}^2\right)}$$

2.2.4. Interpretación.

τ es el decremento relativo en la probabilidad de efectuar una predicción errónea - cuando se pasa del caso en el que no se tiene información al caso en el que se tiene.

El teorema 2.25. señala que la tau simétrica, al igual que τ_j, se puede expresar en términos de la diferencia entre p_{ij} y $p_{i\cdot}.p_{\cdot j}$.

2.2.5. Teorema.

La tau simétrica puede expresarse como

$$\frac{\frac{1}{2}\sum_i\sum_j \frac{(p_{ij}-p_{i\cdot}.p_{\cdot j})^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}}{1-\frac{1}{2}(\sum_i p_{i\cdot}^2+\sum_j p_{\cdot j}^2)}$$

Demostración.

Basta verificar que

$$\sum_i\sum_j \frac{p_{ij}^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}-(\sum_i p_{i\cdot}^2+\sum_j p_{\cdot j}^2)=\sum_i\sum_j \frac{(p_{ij}-p_{i\cdot}.p_{\cdot j})^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}$$

Como $\sum_i\sum_j \frac{(p_{ij}-p_{i\cdot}.p_{\cdot j})^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}$

$$=\sum_i\sum_j \frac{p_{ij}^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}-2\sum_i\sum_j p_{ij}(p_{i\cdot}+p_{\cdot j})+\sum_i\sum_j p_{i\cdot}.p_{\cdot j}(p_{i\cdot}+p_{\cdot j})$$

$$=\sum_i\sum_j \frac{p_{ij}^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}-2\sum_i(p_{i\cdot}\sum_j p_{ij})-2\sum_j(p_{\cdot j}\sum_i p_{ij})+(\sum_j p_{\cdot j})(\sum_i p_{i\cdot}^2)+(\sum_i p_{i\cdot})(\sum_j p_{\cdot j}^2)$$

$$=\sum_i\sum_j \frac{p_{ij}^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}-2\sum_i p_{i\cdot}^2-2\sum_j p_{\cdot j}^2+\sum_i p_{i\cdot}^2+\sum_j p_{\cdot j}^2$$

$$=\sum_i\sum_j \frac{p_{ij}^2(p_{i\cdot}+p_{\cdot j})}{p_{i\cdot}.p_{\cdot j}}-(\sum_i p_{i\cdot}^2+\sum_j p_{\cdot j}^2)$$

Entonces, $\sum_i \sum_j \frac{(p_{ij} - p_{i\cdot}\, p_{\cdot j})^2 (p_{i\cdot} + p_{\cdot j})}{p_{i\cdot}\, p_{\cdot j}} = \sum_i \sum_j \frac{p_{ij}^2 (p_{i\cdot} + p_{\cdot j})}{p_{i\cdot}\, p_{\cdot j}} - \left(\sum_i p_{i\cdot}^2 + \sum_j p_{\cdot j}^2\right)$

por lo tanto, $$\tau = \frac{\frac{1}{2} \sum_i \sum_j \frac{(p_{ij} - p_{i\cdot}\, p_{\cdot j})^2 (p_{i\cdot} + p_{\cdot j})}{p_{i\cdot}\, p_{\cdot j}}}{1 - \frac{1}{2}\left(\sum_i p_{i\cdot}^2 + \sum_j p_{\cdot j}^2\right)}$$

□

Tal y como lo indica el teorema 2.2.6., en una tabla de 2 X 2, los índices τ_j , τ y ρ^2 son numéricamente iguales.

2.2.6. Teorema.

Si las variables A y B son dicotómicas, se cumple que:

$$\tau_j = \tau = \rho^2 = \frac{(p_{11}\, p_{22} - p_{12}\, p_{21})^2}{p_{1\cdot} p_{2\cdot} p_{\cdot 1} p_{\cdot 2}}$$

El corolario que se enuncia a continuación es una consecuencia del teorema que se acaba de mencionar y de otros dos que se demostraron en capítulos previos.

2.2.7. Corolario.

2.2.7.1. En una tabla de 2 X 2, $\tau_j = \tau = \Phi^2$. (Ver el teorema 5.6. del Capítulo II).

2.2.7.2. En una tabla de 2 X 2 estandarizada, $\tau_j = \tau = Y^2$

(Por el teorema 3.6. del Capítulo III).

3. BIBLIOGRAFIA.

1. Bishop, Y.M.M. , Fienberg, S., Holland P.W.:
Discrete Multivariate Analysis: Theory and Practice.
MIT Press, Cambridge Mass., 1975.

2. Goodman, L.A., Kruskal, W.H.:
Measures of association for cross classifications.
Journal of the American Statistical Association,
1954; 49: 732-763.

Measures of association for cross classifications.
II: Further discussion and references.
Journal of the American Statistical Association
1959; 54: 123-163.

Measures of association for cross classifications.
III: Aproximate sampling theory.
Journal of the American Statistical Association.
1963; 58: 310-364.

Measures of association for cross classifications.
IV: Simplification of asymptotic variances.
Journal of the American Statistical Association,
1972; 67: 131-146.

3. Kendall, M.G., Stuart, A.:
The Advanced Theory of Statistics.
C. Griffin and Co. Ltd., London, 2nd. ed.,
Vol. 2: Inference and relationship, 1963.

4. Reynolds, H.T.:
The Analysis of Cross-Classifications.
The Free Press, New York, 1977.

5. Steel, R.G.D., Torrie, J.H.:
Principles and Procedures of Statistics (A Biomedical -
Approach).
McGraw-Hill International Book Co. Singapore, 2nd. ed.,
1980.

V. OTRAS MEDIDAS DE ASOCIACION PARA VARIABLES NOMINALES

1. Medidas de asociación basadas en estructuras latentes.

2. Medidas de disimilaridad.

3. Indices que son iguales a cero si, y solo si, las variables son independientes.

4. Medidas para problemas muy particulares.

5. La asociación entre la variable B y un subconjunto de categorías de A.

6. Bibliografía.

V. OTRAS MEDIDAS DE ASOCIACION

Este Capítulo, a diferencia de los anteriores, es muy heterogéneo en su contenido porque trata de medidas muy distintas entre si; además, únicamente se enuncian los índices poblacionales.

En primer lugar se explicará qué es una estructura latente y como obtener medidas de asociación a partir de ella. Después se verán algunas medidas de disimilaridad que sirven para determinar qué tan distintas son dos poblaciones respecto a una cierta variable. En la tercera parte de este Capítulo se enuncian algunas medidas de asociación que tienen la propiedad de ser iguales o cero únicamente cuando las variables son independientes. Posteriormente, en la cuarta sección se verán algunos índices propuestos para tres problemas muy particulares, dos de ellos meteorológicos y el otro antropológico. Por último, en la quinta parte del Capítulo se explica como medir la asociación entre una variable y un subconjunto de categorías de otra variable.

1. MEDIDAS DE ASOCIACION BASADAS EN ESTRUCTURAS LATENTES.

En ocasiones una tabla de contingencia es considerada como un promedio o una ponderación de dos o más tablas. A estas tablas se les conoce como estructuras latentes y tienen en común algunas características; por ejemplo, el ser mutuamente independientes.

Las medidas de asociación son, en este contexto un promedio o una ponderación de una característica numérica de las estructuras latentes y deben ser expresadas en términos de las probabilidades de la tabla original.

1.1. La Probabilidad de Exito de un Pronóstico.

La medida de asociación que se verá a continuación fue propuesta de manera independiente por Pierce y Youden con el fin de saber que tan confiables eran los pronósticos de un meteorólogo.

1. Planteamiento del Problema.

Una persona, basada en diversos estudios sobre el ambiente pronostica varias veces y de manera independiente, si habrá o no un tornado. Una vez hecho el pronóstico, se deja transcurrir cierto tiempo antes de determinar si fue correcto o incorrecto. Con base en la tabla 1.1. se desea calcular:

1.1. La proporción de pronósticos correctos, sin tomar en cuenta el azar y

1.2. La probabilidad de que por azar se pronostique que habrá un tornado.

Tabla 1.1.

Pronóstico	Tornado	No Tornado
Tornado	p_{11}	p_{12}
No Tornado	p_{21}	p_{22}
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$

2. Solución.

Para resolver este problema, Pierce y Youden consideraron que dos personas realizaron los pronósticos de manera independiente. Una de ellas era infalible; la otra usaba un procedimiento aleatorio para hacer los pronósticos. Con este argumento, Pierce y Youden, descompusieron la tabla anterior en dos tablas; una para cada persona. Se denotará como θ a la proporción de pronósticos realizados por el pronosticador infalible y su complemento, $1-\theta$, la proporción de pronósticos hechos por la otra persona. Además γ será la probabilidad de que la persona que hace los pronósticos aleatoriamente diga que ocurrirá un tornado y $1-\gamma$ será la probabilidad de que esta misma persona diga que no ocurrirá. Entonces, la tabla de contingencia para el pronosticador infalible es:

	Tornado	No Tornado
Tornado	$\theta p_{\cdot 1}$	0
No Tornado	0	$\theta p_{\cdot 2}$

y para la otra persona es:

	Tornado	No Tornado
Tornado	$(1-\theta)\gamma\ p_{\cdot 1}$	$(1-\theta)\gamma\ p_{\cdot 2}$
No Tornado	$(1-\theta)(1-\gamma)p_{\cdot 1}$	$(1-\theta)(1-\gamma)p_{\cdot 2}$

Por lo tanto, la tabla 1.1. se puede reexpresar como:

	Tornado	No Tornado
Tornado	$p_{11} = \theta\, p_{\cdot 1} + (1-\theta)\, p_{\cdot 1}\psi$	$p_{12} = (1-\theta)\,\psi\; p_{\cdot 2}$
No Tornado	$p_{21} = (1-\theta)(1-\psi)\, p_{\cdot 1}$	$p_{22} = \theta\, p_{\cdot 2} + (2-\theta)(1-\psi)\, p_{\cdot 2}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$

El objetivo es expresar a θ y a ψ en términos de p_{ij}, $p_{i\cdot}$ y $p_{\cdot j}$ (con i=1, 2 y j= 1, 2).

De la última tabla de contingencia se tiene que

$$p_{11} = \theta\, p_{\cdot 1} + (1-\theta)\, p_{\cdot 1}\psi$$

y $p_{12} = (1-\theta)\,\psi\; p_{\cdot 2} \Longrightarrow \dfrac{p_{12}}{p_{\cdot 2}} = (1-\theta)\psi$ ... (E-5.1.1.)

Sustituyendo $(1-\theta)\psi$ en la primera ecuación:

$$p_{11} = \theta\, p_{\cdot 1} + \frac{p_{\cdot 1}\, p_{12}}{p_{\cdot 2}}$$

$$\Longleftrightarrow \theta = \frac{p_{11}}{p_{\cdot 1}} - \frac{p_{12}}{p_{\cdot 2}}$$

El valor de ψ se obtiene a partir de la ecuación (E-5.1.1.), es decir,

$$\psi = \frac{p_{12}}{p_{\cdot 2}\,(1-\theta)}$$

en donde, $p_{\cdot 2}(1-\theta) = p_{\cdot 2}\left(1-\frac{p_{11}}{p_{\cdot 1}}+\frac{p_{12}}{p_{\cdot 2}}\right) = p_{\cdot 2}\left(\frac{p_{\cdot 1}p_{\cdot 2}-p_{11}\,p_{\cdot 2}+p_{12}\,p_{\cdot 1}}{p_{\cdot 1}\;p_{\cdot 2}}\right)$

$$= \frac{p_{12}\,p_{\cdot 1}+p_{\cdot 2}\,(p_{\cdot 1}-p_{11})}{p_{\cdot 1}} = \frac{p_{12}\,p_{\cdot 1}+p_{\cdot 2}\,p_{21}}{p_{\cdot 1}}$$

entonces, $$\psi = \frac{p_{12}\,p_{\cdot 1}}{p_{12}\,p_{\cdot 1} + p_{21}\,p_{\cdot 2}}$$

Por lo tanto, las medidas propuestas por Pierce y Youden son las siguientes.

3. Medidas Propuestas.

3.1. El índice que mide la proporción de pronósticos correctos, sin tomar en cuenta el azar es:

$$\theta = \frac{p_{11}}{p_{\cdot 1}} - \frac{p_{12}}{p_{\cdot 2}}$$

3.2. La probabilidad de que, por azar, se pronostique un tornado es :

$$\psi = \frac{p_{\cdot 1}\,p_{12}}{p_{\cdot 1}\,p_{12} + p_{\cdot 2}\,p_{21}}$$

4. Interpretación.

El índice θ compara qué tan grande es la probabilidad condicional de predecir un tornado dado que éste va a ocurrir con la probabilidad condicional de predecir un tornado dado que éste no va a ocurrir. Si

4.1 $\Theta = 0$ entonces la probabilidad condicional de predecir un tornado dado que éste va a ocurrir es igual a la probabilidad condicional, de predecir un tornado dado a - que éste no va a ocurrir. Es decir, independientemente de la información que tenga el meteorólogo de las condiciones ambientales, su pronóstico siempre es el mismo. En - otras palabras, $\Theta = 0$ si las variables son independientes.

4.2. $\Theta = 1$ si todos los pronósticos fueron correctos y ninguno de ellos fue aleatorio.

1.2. Una Medida Propuesta por Lazarsfeld y Kendall.

El índice propuesto por Lazarsfeld y Kendall mide - cuál es la probabilidad de que una persona no sea honesta al contestar una pregunta.

1. Planteamiento del Problema.

A una población se le hace una misma pregunta en dos ocasiones distintas. Se supondrá que:

1.1. La pregunta solo puede ser contestada afirmativa o - negativamente.

1.2. Las dos respuestas deben ser iguales.

1.3. Una o ambas respuestas pueden ser deshonestas.(*)

1.4. La probabilidad de que una respuesta sea deshonesta es pequeña (menor o igual a 0.5).

(*) Se dirá que una respuesta es "deshonesta" si el individuo no entendió la pregunta, estaba influenciado por otras - opiniones, etc.

y 1.5. La tabla de contingencia obtenida es:

Tabla 1.2.

Primera Respuesta	Segunda Respuesta: Si	No	Totales
Si	p_{11}	p_{12}	$p_{1\cdot}$
No	p_{21}	p_{22}	$p_{2\cdot}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	1

El objetivo es definir, a partir de esta tabla, un índice que mida la probabilidad de que una de las respuestas dadas - por una misma persona sea deshonesta.

2. Solución.

Lazarsfeld y Kendall descompusieron la tabla 1.2. en - dos tablas; una para los individuos que debieron haber contestado "Si" las dos veces y otra para los que debieron -- haber contestado "No" en las dos ocasiones. La proporción de individuos en la primera tabla se denotará como K_1 y - su complemento, $1-K_1 = K_2$, será la proporción de individuos en la segunda tabla. Además, α denotará la probabilidad de que alguien haya contestado "No" en lugar de "Si". Por lo tanto, $1-\alpha$ es la probabilidad de que alguien haya contestado "Si" honestamente; "y" denotará la probabilidad de - que alguien haya contestado "Si" en lugar de "No" y $1-y$ es la probabilidad de que alguien haya contestado "No" -

honestamente.

Si las respuestas de quienes tenían que contestar " Si" - son independientes de las respuestas de quienes tenían que contestar "No", entonces, la tabla 1.2. es igual a la suma de las dos tablas siguientes:

	Si	No
Si	$K_1(1-x)^2$	$K_1(1-x)x$
No	$K_1(1-x)x$	K_1x^2

	Si	No
Si	K_2y^2	$K_2y(1-y)$
No	$K_2(1-y)y$	$K_2(1-y)^2$

Es decir, la tabla 1.2. se puede reexpresar de la manera siguiente:

	Si	No	Totales
Si	$p_{11}=K_1(1-x)^2+K_2y^2$	$p_{12}=K_1(1-x)x+K_2y(1-y)$	$p_{1\cdot}=K_1(1-x)+K_2y$
No	$p_{21}=K_1x(1-x)+K_2(1-y)y$	$p_{22}=K_1x^2+K_2(1-y)^2$	$p_{2\cdot}=K_1x+K_2(1-y)$
Totales	$p_{\cdot1}=K_1(1-x)+K_2y$	$p_{\cdot2}=K_1x+K_2(1-y)$	1

En esta tabla, $p_{12}=p_{21}$ por lo tanto, $p_{\cdot1}=p_{1\cdot}$ y $p_{\cdot2}=p_{2\cdot}$. Además, de las tres ecuaciones:

$$p_{11} = K_1(1-x)^2 + K_2y^2$$

$$p_{12} = K_1(1-x)x + K_2(1-y)y$$

$$p_{22} = K_1x^2 + K_2(1-y)^2$$

solo dos son linealmente independientes, por lo tanto, los parámetros (K_1, x y y) no se pueden expresar, de manera única como función de las probabilidades poblacionales. Por esto, Lazarsfeld y Kendall tuvieron que hacer el supuesto adicional de que x=y.

Sustituyendo en la tabla anterior "y" por "x" y K_2 por $1-K_1$ se obtiene la tabla:

	Si	No	Totales
Si	$p_{11}=x^2-2K_1x+K_1$	$p_{12}=x(1-x)$	$p_{1\cdot}=K_1(1-2x)+x$
No	$p_{21}=x(1-x)$	$p_{22}=x^2-2(1-K_1)x+(1-K_1)$	$p_{2\cdot}=K_1(1-2x)+1-x$
Totales	$p_{\cdot 1}=K_1(1-2x)+x$	$p_{\cdot 2}=-K_1(1-2x)+1-x$	1

Para calcular x, la probabilidad de que un individuo sea deshonesto, se resuelve la ecuación $p_{12}=x(1-x)$. Las soluciones son:

$$x_1 = \frac{1}{2}\left(1 + \sqrt{1-4p_{12}}\right)$$

$$x_2 = \frac{1}{2}\left(1 - \sqrt{1-4p_{12}}\right)$$

Como por hipótesis, la probabilidad de que una respuesta sea deshonesta es menor o igual a 0.5 se concluye que $x = x_2$.

En resumen, haciendo los supuestos de que $x = y$ y $x \leq \frac{1}{2}$ se obtiene la siguiente medida de la deshonestidad.

3. Medida Propuesta.

La probabilidad de que una persona responda alguna pregunta deshonestamente es:

$$x = \frac{1}{2}\left(1 - \sqrt{1 - 4p_{12}}\right)$$

4. Observaciones.

4.1. Conociendo el valor de x se puede calcular:

4.1.1. La probabilidad de que una respuesta sea honesta y la otra deshonesta, es decir $2x\,(1-x)$.

4.1.2. La probabilidad de que ambas respuestas sean honestas o de que ambas sean deshonestas, es decir $1 - 2x\,(1-x)$.

4.1.3. K_1, la proporción de individuos que debieron haber contestado "Si" a las dos preguntas. (Para ello, se despeja K_1 de la ecuación $p_{1\cdot} = K_1(1-2x)+x$ y se sustituye x. De esta manera se obtiene que

$$K_1 = \frac{2p_{1\cdot} - 1}{2\sqrt{1 - 4p_{12}}} + \frac{1}{2})$$

4.2. Si la primera respuesta es siempre independiente de la segunda, entonces $K_1 = 0$ ó $K_1 = 1$ ó $x = \frac{1}{2}$

Demostración.

Si las dos respuestas son independientes, entonces $p_{ij} = p_{i\cdot}p_{\cdot j}$ $\forall$ i,j.; en particular, $p_{11} = p_{1\cdot}p_{\cdot 1}$. Esto quiere decir en términos de la última tabla, que:

$$x^2 - 2K_1 x + K_1 = \left[K_1(1-2x)+x\right]^2$$

$$\Leftrightarrow \quad x^2 - 2K_1 x + K_1 = K_1^2(1-2x)^2 + 2xK_1(1-2x) + x^2$$

$$\Leftrightarrow \quad K_1(1-2x) = K_1^2(1-2x)^2 + 2xK_1(1-2x)$$

$$\Leftrightarrow \quad 0 = K_1^2(1-2x)^2 + 2xK_1(1-2x) - K_1(1-2x)$$

$$\Leftrightarrow \quad 0 = K_1^2(1-2x)^2 - K_1(1-2x)^2$$

$$\Leftrightarrow \quad 0 = K_1(1-2x)^2(K-1)$$

$$\Leftrightarrow \quad K_1 = 0 \text{ ó } K_1 = 1 \text{ ó } x = \frac{1}{2}$$

Análogamente se demuestra que $p_{12} = p_{21} = p_{2\cdot}p_{\cdot 1}$ y que $p_{22} = p_{2\cdot}p_{\cdot 2}$.

5. Desventajas.

El modelo anterior es muy restrictivo, pocas tablas de contingencia cumplen con todos los supuestos que se mencionan.

1.3. Una Medida Propuesta por Goodman y Kruskal.

Goodman y Kruskal plantearon un problema similar al de la sección anterior y el modelo que usaron para resolverlo es parecido al de Lazarsfeld y Kendall solo que es menos restrictivo y además permite calcular dos medidas de deshonestidad. Bajo ciertas condiciones, una de estas medidas es igual a la de Lazarsfeld y Kendall.

1. Planteamiento del Problema.

A una población se le hicieron dos preguntas distintas

pero íntimamente relacionadas. Las respuestas solo podían ser "Si" o "No" y la tabla que se obtuvo fue la siguiente:

Tabla 1.3.

Respuesta 1: \ Respuesta 2:	Si	No	Totales
Si	p_{11}	p_{12}	$p_{1\cdot}$
No	p_{21}	p_{22}	$p_{2\cdot}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	1

Con esta información se desea saber:

1.1. Cuál es la probabilidad de que la respuesta 1 haya sido deshonesta y

1.2. Cuál es la probabilidad de que la respuesta 2 haya sido deshonesta.

2. Solución.

Goodman y Kruskal partieron del supuesto de que la población estudiada se podía dividir en dos grupos, uno formado por una proporción de K_1 individuos que debieron haber contestado "Si" a las dos preguntas y otro grupo constituido por una proporción de $K_2 = 1 - K_1$ individuos cuyas dos respuestas debieron haber sido "No".

La probabilidad de que una persona contestara deshonestamente las preguntas 1 y 2 se denotará, respectivamente, como x_1 y x_2; se supondrá que x_1 y x_2 no dependen

del grupo al que pertenecen los individuos. La tabla de contingencia para los que debieron haber contestado "Si" a las dos preguntas es:

	Respuesta 2: Si	Respuesta 2: No
Respuesta 1: Si	$K_1(1-x_1)(1-x_2)$	$K_1(1-x_1)x_2$
Respuesta 1: No	$K_1x_1(1-x_2)$	$K_1x_1x_2$

y la tabla para las que debieron haber contestado "No" a las dos preguntas es:

	Respuesta 2: Si	Respuesta 2: No
Respuesta 1: Si	$K_2x_1x_2$	$K_2x_1(1-x_2)$
Respuesta 1: No	$K_2(1-x_1)x_2$	$K_2(1-x_1)(1-x_2)$

Por lo tanto, la tabla 1.3. se puede reexpresar como:

	Respuesta 2: Si	Respuesta 2: No	Totales
Respuesta 1: Si	$K_1(1-x_1)(1-x_2)+K_2x_1x_2$	$K_1(1-x_1)x_2+K_2x_1(1-x_2)$	$p_{1\cdot}=K_1(1-x_1)+K_2x_1$
Respuesta 1: No	$K_1x_1(1-x_2)+K_2(1-x_1)x_2$	$K_1x_1x_2+K_2(1-x_1)(1-x_2)$	$p_{2\cdot}=K_1x_1+K_2(1-x_1)$
Totales	$p_{\cdot1}=K_1(1-x_2)+K_2x_2$	$p_{\cdot2}=K_1x_2+K_2(1-x_2)$	1

El objetivo es expresar a x_1 y a x_2 como función de las probabilidades p_{ij}. Despejando x_1 de la ecuación $p_{1\cdot} = k_1 (1 - x_1) + K_2 x_1$ se obtiene que: $x_1 = \frac{p_{1\cdot} - K_1}{1 - 2K_1}$ y despejando x_2 de la ecuación $p_{\cdot 1} = K_1(1-x_2) + K_2 x_2$ se obtiene que - $x_2 = \frac{p_{\cdot 1} - K_1}{1 - 2K_1}$. Ahora, para determinar el valor de x_1 y x_2 hay que calcular K_1 y para ello hay que demostrar que - $\frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{1 - 2(p_{12}+p_{21})} = K_1(1-K_1)$. Si $\frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{1 - 2(p_{12}+p_{21})}$ se denota - como R entonces, $K_1^2 - K_1 + R = 0$. Las soluciones de esta ecuación son $K_1 = \frac{1}{2}\left(1 \pm \sqrt{1 - 4R}\right)$ en donde $0 \leq 1 - 4R \leq 1$ si, y solo si, $0 \leq R \leq \frac{1}{4}$. Sustituyendo K_1 en x_1 y x_2 se - obtienen las dos medidas siguientes.

3. Medidas Propuestas.

La probabilidad de que una persona haya contestado deshonestamente la pregunta 1 es:

$$x_1 = \frac{1}{2} \pm \frac{2p_{1\cdot} - 1}{2\sqrt{1 - 4R}}$$

y la probabilidad de que alguien haya contestado deshonestamente la pregunta 2 es:

$$x_2 = \frac{1}{2} \pm \frac{2p_{\cdot 1} - 1}{2\sqrt{1 - 4R}} \quad \text{con} \quad R = \frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{1-2(p_{12}+p_{21})} \in \left[0, \frac{1}{4}\right)$$

4. Comentarios.

4.1. Los índices x_1 y x_2 son menores o iguales a un -

medio cuando $\left(p_{1\cdot} - \frac{1}{2}\right)\left(p_{\cdot 1} - \frac{1}{2}\right) \geq 0$.

4.2. x_1 y x_2 pueden promediarse para obtener una sola medida de deshonestidad. En este caso, mediante un procedimiento aleatorio (lanzar una moneda) se puede decidir qué respuesta fue deshonesta.

4.3. $x_1 x_2 + (1-x_1)(1-x_2)$ es la probabilidad de que una persona, escogida al azar, haya contestado ambas preguntas honesta o deshonestamente.

4.4. Las dos respuestas son independientes si, y solo si, $k_1 = 0$ ó $k_1 = 1$ ó $x_1 = \frac{1}{2}$ ó $x_2 = \frac{1}{2}$.

4.5. Cuando $p_{12} = p_{21}$, $x_1 = x_2 = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\,p_{12}}\right)$; si se escoge la raíz negativa se obtiene el índice de Lazarsfeld y Kendall (ver la sección anterior).

2. MEDIDAS DE DISIMILARIDAD.

En las ciencias sociales a veces se clasifica a dos poblaciones (Hombres-Mujeres o Negros-Blancos) de acuerdo a la fecha o al área geográfica en la que fueron estudiadas. La tabla resultante tiene dimensión I X 2 si las columnas se refieren a las poblaciones de interés y los renglones a la variable tiempo o lugar. En este caso, usando un índice de disimilaridad se puede medir qué tan distintas son las dos poblaciones al ir comparándolas renglón por renglón.

A continuación se analizan dos medidas de disimilaridad para tablas de I X 2. Después se explica como calcular una medida de disimilaridad en una tabla de I X J en donde J denota el número de poblaciones en estudio.

2.1. Indice de Gini, Florence et al.

Para medir en una tabla de I X 2, que tan distintas o disímiles son dos poblaciones respecto a una variable A se puede calcular el índice que se define a continuación.

1. Definición.

Una medida de disimilaridad que ha sido propuesta por Gini, Florence, Hoover, Duncan y Duncan, y Bogue es:

$$D = \frac{1}{2} \sum_{i=1}^{I} \left| \frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right|$$

D es igual a un medio de la suma de las diferencias, en valor absoluto, de las probabilidades condicionales de la primera y la segunda columnas.

2. Otras Expresiones para D.

2.1. Las diferencias $\left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right)$ pueden ser positivas o negativas. Si son positivas, entonces

$$\left| \frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right| = \frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}}$$

y si son negativas, entonces

$$\left| \frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right| = - \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right)$$

Por lo tanto,

$$D = \frac{1}{2} \sum_i \left| \frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right| = \frac{1}{2} \left[\sum_i^{+} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right) + \sum_i^{-} - \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right) \right]$$

en donde

$\sum_i^{+}$ es la suma de todas las diferencias positivas y

$\sum_i^{-}$ es la suma de todas las diferencias negativas.

Por otro lado,

$\sum_i \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right) = 1 - 1 = 0$. Esto quiere decir que

$$\sum_i^{+} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right) = \sum_i^{-} - \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right)$$

De aquí que:

$$D = \sum_i^{+} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right) = - \sum_i^{-} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right)$$

Como $\sum_i^{+} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right)$ es la suma de todas aquellas diferencias en las que, la probabilidad condicional del i-ésimo renglón dada la columna uno es mayor que la probabilidad condicional del i-ésimo renglón dada la columna 2, entonces, D se puede interpretar como la proporción mínima de individuos de la columna 1 -

que deben ser cambiados de renglón con el fin de que, para toda i, $\frac{p_{i1}}{p_{\cdot 1}}$ sea igual a $\frac{p_{i2}}{p_{\cdot 2}}$.

Análogamente, la expresión $-\sum_i^- \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}}\right)$ - indica cuál es la mínima proporción de individuos, - pertenecientes a la columna 2, que deben ser cambiados de renglón para que $\frac{p_{i1}}{p_{\cdot 1}} = \frac{p_{i2}}{p_{\cdot 2}}$.

2.2. De acuerdo con Goodman y Kruskal, en algunos - contextos puede ser de gran interés reacomodar los - renglones de la tabla de acuerdo a la magnitud de - las diferencias entre $\frac{p_{i1}}{p_{\cdot 1}}$ y $\frac{p_{i2}}{p_{\cdot 2}}$. Hasta arriba - se puede poner al renglón con la diferencia más grande, la seguiría el renglón con la segunda diferencia más grande, etc. Si se tienen I_* renglones para los cuales $\frac{p_{i1}}{p_{\cdot 1}} \geq \frac{p_{i2}}{p_{\cdot 2}}$, entonces, en términos de la tabla reordenada,

$$D = \sum_{i=1}^{I_*} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}}\right) = -\sum_{i=I_*+1}^{I} \left(\frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}}\right)$$

2.3. Otras expresiones para D son:

$$D = \sum_i \frac{1}{2p_{\cdot 2}} \left| \frac{p_{i1}}{p_{\cdot 1}} - p_{i\cdot} \right|$$

$$= \sum_i \frac{1}{2p_{\cdot 1}} \left| \frac{p_{i2}}{p_{\cdot 2}} - p_{i\cdot} \right|$$

$$D = \sum_{i=1}^{I}\sum_{j=1}^{2} \left| \frac{p_{ij}}{p_{\cdot j}} - p_{i\cdot} \right| \frac{1}{4(1-p_{\cdot j})}$$

$$= \sum_{i} \left| \frac{p_{i1}}{p_{i\cdot}} - p_{\cdot 1} \right| \frac{p_{i\cdot}}{2 p_{\cdot 1} p_{\cdot 2}}$$

$$= \sum_{i} \left| \frac{p_{i2}}{p_{i\cdot}} - p_{\cdot 2} \right| \frac{p_{i\cdot}}{2 p_{\cdot 1} p_{\cdot 2}}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{2} \left| \frac{p_{ij}}{p_{i\cdot}} - p_{j} \right| \frac{p_{i\cdot}}{4 p_{\cdot 1} p_{\cdot 2}}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{2} \frac{|p_{ij} - p_{i\cdot} p_{\cdot j}|}{4 p_{\cdot 1} p_{\cdot 2}}$$

La última fórmula es parecida a la del coeficiente de contingencia en media cuadrática, solo que en Φ^2, las diferencias ($p_{ij} - p_{i\cdot}.p_{\cdot j}$) están elevadas al cuadrado y ponderadas de otra manera. A continuación se demuestran algunas de las expresiones anteriores.

Demostración.

2.3.1. Por demostrar que $D = \sum_{i} \frac{1}{2 p_{\cdot 2}} \left| \frac{p_{i1}}{p_{\cdot 1}} - p_{i\cdot} \right|$

Como $$D = \frac{1}{2} \sum_{i} \left| \frac{p_{i1}}{p_{\cdot 1}} - \frac{p_{i2}}{p_{\cdot 2}} \right|$$

$$\therefore D = \sum_{i} \frac{1}{2} \left| \frac{p_{i1} p_{\cdot 2}}{p_{\cdot 1} p_{\cdot 2}} - \frac{p_{i2}}{p_{\cdot 2}} \right| = \sum_{i} \frac{1}{2} \left| \frac{p_{i1}(1-p_{\cdot 1})}{p_{\cdot 1} p_{\cdot 2}} - \frac{p_{i2}}{p_{\cdot 2}} \right|$$

$$= \sum_{i} \frac{1}{2} \left| \frac{p_{i1}}{p_{\cdot 1} p_{\cdot 2}} - \frac{p_{i1}}{p_{\cdot 2}} - \frac{p_{i2}}{p_{\cdot 2}} \right| = \sum_{i} \frac{1}{2} \left| \frac{p_{i1}}{p_{\cdot 1} p_{\cdot 2}} - \frac{p_{i\cdot}}{p_{\cdot 2}} \right|$$

$$= \sum_{i} \frac{1}{2 p_{\cdot 2}} \left| \frac{p_{i1}}{p_{\cdot 1}} - p_{i\cdot} \right|$$

2.3.3. Por demostrar que $D = \sum_i \left| \frac{p_{i1}}{p_{i.}} - p_{.1} \right| \frac{p_{i.}}{2 p_{.1} p_{.2}}$

De acuerdo con la demostración 1:

$$D = \sum_i \frac{1}{2 p_{.2}} \left| \frac{p_{i1}}{p_{.1}} - p_{i.} \right|$$

además,

$$\sum_i \frac{1}{2 p_{.2}} \left| \frac{p_{i1}}{p_{.1}} - p_{i.} \right| = \sum_i \left| \frac{p_{i1} p_{i.} - p_{.1} p_{i.}}{p_{i.} p_{.1}} \right| \frac{1}{2 p_{.2}}$$

$$\Rightarrow D = \sum_i \frac{1}{2 p_{.2}} \left| \frac{p_{i1}}{p_{.1}} - p_{i.} \right| = \sum_i \left| \frac{p_{i1}}{p_{i.}} - p_{.1} \right| \frac{p_{i.}}{2 p_{.1} p_{.2}}$$

Por lo tanto, $D = \sum_i \left| \frac{p_{i1}}{p_{i.}} - p_{.1} \right| \frac{p_{i.}}{2 p_{.1} p_{.2}}$

Análogamente se demuestra que $D = \sum_i \left| \frac{p_{i2}}{p_{i.}} - p_{.2} \right| \frac{p_{i.}}{2 p_{.1} p_{.2}}$

2.3.4. Por demostrar que $D = \sum_i \sum_{j=1}^{2} \frac{| p_{ij} - p_{i.} p_{.j} |}{4 p_{.1} p_{.2}}$

De acuerdo con la demostración 2,

$$D = \sum_i \sum_{j=1}^{2} \left| \frac{p_{ij}}{p_{.j}} - p_{i.} \right| \frac{1}{4(1-p_{.j})}$$

y como $\sum_i \sum_{j=1}^{2} \left| \frac{p_{ij}}{p_{.j}} - p_{i.} \right| \frac{1}{4(1-p_{.j})} = \sum_i \sum_{j=1}^{2} \left| \frac{p_{ij} - p_{i.} p_{.j}}{p_{.j}} \right| \frac{1}{4(1-p_{.j})}$

$$= \sum_i \left\{ \frac{| p_{i1} - p_{i.} p_{.1} |}{4 p_{.1} (1-p_{.1})} + \frac{| p_{i2} - p_{i.} p_{.2} |}{4 (1-p_{.2}) p_{.2}} \right\}$$

$$= \sum_i \sum_{j=1}^{2} \frac{| p_{ij} - p_{i.} p_{.j} |}{4 p_{.1} p_{.2}}$$

Entonces, $D = \sum_i \sum_{j=1}^{2} \left| \frac{p_{ij}}{p_{.j}} - p_{i.} \right| \frac{1}{4(1-p_{.j})} = \sum_i \sum_{j=1}^{2} \frac{| p_{ij} - p_{i.} p_{.j} |}{4 p_{.1} p_{.2}}$

$$\Rightarrow D = \sum_i \sum_{j=1}^{2} \frac{| p_{ij} - p_{i.} p_{.j} |}{4 p_{.1} p_{.2}}$$

2.2. Medidas de disimilaridad para tablas de I X J.

A partir de una tabla de I X J se puede medir al grado de disimilaridad entre J poblaciones de la manera siguiente:

1. Se descompone la tabla de I X J en varias subtablas de I X 2. Las columnas de estas subtablas pueden ser todas las parejas de clases de B que se puedan formar a partir de las J categorías de la tabla original. De esta forma se obtienen $\frac{J(J-1)}{2}$ tablas de I X 2. Otra manera de dividir una tabla de I X J en varias subtablas de I X 2 es considerar que una de las columnas de las subtablas se refiere a una categoría de B y la otra a las frecuencias marginales de A. Al proceder de esta manera se obtienen J tablas de I X 2.
2. Después, para cada subtabla (de I X 2) se calcula alguna medida de disimilaridad.
3. Los índices calculados se promedian para obtener una medida global.

3. INDICES QUE SON IGUALES A CERO SI, Y SOLO SI, LAS VARIABLES SON INDEPENDIENTES.

Las tres medidas de asociación que se verán a continuación tienen la propiedad de ser iguales a cero si, y solo si, las variables son independientes.

3.1. En 1924, H. Cramér propuso la siguiente medida para tablas de I X J:

$$\text{mín} \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - u_i v_j)^2$$

el mínimo se calcula sobre todos los números $u_1, u_2, \ldots, u_I, v_1, v_2, \ldots v_J$.

Algunas propiedades de este índice son:

3.1.1. Es igual a cero si, y solo si, las variables son independientes.

3.1.2. Siempre es menor o igual a 0.25.

y 3.1.3. No toma ningún valor en particular cuando la asociación entre las variables es perfecta.

3.2. Con el fin de medir la asociación entre dos variables con I y J categorías, J. F. Steffensen definió en 1933:

$$\psi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \frac{(p_{ij} - p_{i\cdot}\, p_{\cdot j})^2}{p_{i\cdot}\,(1-p_{i\cdot})\, p_{\cdot j}\,(1-p_{\cdot j})}$$

El método que utilizó para construir este índice es similar al que usó Jordan para generalizar el cociente de productos cruzados (ver la Sección 6.2. del Capítulo III). Primero descompuso una tabla de I X J en I.J subtablas de 2 X 2 tomando cada una de las celdas de la tabla original y su complemento. Después, para cada tabla de 2 X 2 calculó el coefi-

ciente de contingencia en media cuadrática y los promedió.

Otra medida propuesta por Steffensen es:

$$\omega = \frac{2 \sum\sum (p_{ij} - p_{i\cdot}p_{\cdot j})}{\sum\sum (p_{ij} - p_{i\cdot}p_{\cdot j}) + 1 - \sum\sum p_{ij}^2}$$

en donde $\sum\sum$ es la suma sobre todas las celdas para las cuales $p_{ij} > p_{i\cdot}p_{\cdot j}$.

φ^2 y ω tienen las propiedades siguientes:

3.2.1. Toman valores en el intervalo $[0,1]$

3.2.2. Son iguales a cero si, y solo si, las variables son independientes.

y 3.2.3. Son iguales a uno cuando la asociación es perfecta estricta.

3.3. Otra medida de asociación que es igual a cero si, y solo si, las variables son independientes es $\frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}$. Este es un índice propuesto por H. Eyraud para tablas de 2 x 2. (Si el numerador estuviera elevado al cuadrado esta medida de asociación sería igual a ρ^2 ya que $p_{11} - p_{1\cdot}p_{\cdot 1} = p_{11}p_{22} - p_{12}p_{21}$).

4. MEDIDAS PARA PROBLEMAS MUY PARTICULARES.

4.1. Un Problema Meteorológico.

En 1870, W. Köppen se interesó por medir la constan-

cia de un fenómeno meteorológico a través del tiempo. Köppen trabajó con tablas de 2 X 2 como la siguiente:

Dirección del viento a las 2 p.m. en la estación de observación

Dirección del viento antes de las 8 a.m. en la estación de observación.	Norte	No-Norte	Totales
Norte	p_{NN}	$p_{N\bar{N}}$	p_N
No-Norte	$p_{\bar{N}N} = p_{N\bar{N}}$	$p_{\bar{N}\bar{N}}$	$1-p_N$
Totales	p_N	$1-p_N$	1

Se puede observar que las categorías y las frecuencias marginales de las dos variables son iguales.

Para medir, con base en la tabla anterior, qué tan constante es la dirección del viento Norte y No-Norte entre las 8 a.m. y las 2 p.m. Köppen propuso:

$\frac{p_N(1-p_N) - p_{N\bar{N}}}{p_N(1-p_N)}$. Esta medida es igual a cero si la dirección del viento es independiente del tiempo.

4.2. Un Problema Antropológico.

En 1932, H.E. Driver y A. L. Kroeber propusieron tres índices para determinar qué tan relacionadas están dos sociedades que presentan algunas características comunes y otras no.

Las tablas con las que trabajaron Driver y Kroeber - son como la siguiente:

		Sociedad B		
		Características Comunes	Características Diferentes	Totales
Sociedad A	Características Comunes	p_{11}	p_{12}	$p_{1\cdot}$
	Características Diferentes	p_{21}	p_{22}	$p_{2\cdot}$
	Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	1

en donde, por ejemplo, p_{21} denota la proporción de características que se encontraron en la sociedad B y que no tuvo la sociedad A.

Las medidas que propusieron son:

$$\frac{p_{11}}{2}\left(\frac{1}{p_{1\cdot}} + \frac{1}{p_{\cdot 1}}\right), \quad \frac{p_{11}}{\sqrt{p_{1\cdot}p_{\cdot 1}}}, \quad \frac{p_{11}}{1 - p_{22}}$$

4.3. Otro Problema Meteorológico.

El problema de medir qué tan confiables son los pronósticos meteorológicos ha sido estudiado por varias personas. Cuando se expusieron las medidas de asociación basadas en estructuras latentes se explicó - como fue resuelto este problema por Pierce y Youden. Las tablas de contingencias con las que otras personas trabajaron son similares a la tabla 5.1.1. Es -

decir, dichas tablas son de 2 X 2 y una de las variables se refiere al pronóstico realizado por el meteorólogo (ocurrirá o no ocurrirá el fenómeno natural) y la otra variable se refiere a lo que sucedió en la naturaleza (ocurrió o no ocurrió el fenómeno natural):

		B: Naturaleza	
		Ocurrió	No ocurrió
A: Meteorólogo	Ocurrirá	p_{11}	p_{12}
	No ocurrirá	p_{21}	p_{22}

Para medir el porcentaje de pronósticos correctos Finley propuso en 1884: $\left(\sum_i p_{ii}\right) 100\%$. El defecto de esta medida es que toma en cuenta los pronósticos que por azar fueron acertados y en realidad lo que se quiere saber es si el meteorólogo tiene alguna habilidad para predecir si un fenómeno va a ocurrir o no. Gilbert, en desacuerdo con el índice de Finley propuso:

$$\frac{p_{11} - p_{1\cdot}p_{\cdot 1}}{p_{1\cdot} + p_{\cdot 1} - p_{11} - p_{1\cdot}p_{\cdot 1}}$$

Esta medida siempre es menor o igual a uno y toma el valor de cero cuando las variables son independientes, es decir, cuando el meteorólogo realiza los pronósticos ignorando las condiciones ambientales.

5. LA ASOCIACION ENTRE B Y UN SUBCONJUNTO DE CATEGORIAS DE A.

Cuando la asociación entre las variables A y B es débil puede ser interesante medir la asociación que existe, por ejemplo, entre la variable B y un subconjunto de categorías de A.

Como se verá, los procedimientos propuestos para medir la asociación en este tipo de problemas varían de acuerdo a los objetivos del estudio. Se partirá del supuesto de que A tiene I clases y que B tiene J.

5.1. La Asociación entre B y una Categoría de A.

Para medir la asociación en tre B y la clase A_{i_o} se reduce la tabla de I X J a una tabla de 2 X J en la que un renglón se refiere a la categoría A_{i_o} y el otro a todas las categorías de A distintas de A_{i_o}. Es decir, la tabla que se obtiene es:

	B_1	B_2		B_J
A_{i_o}	p_{i_o1}	p_{i_o2}		p_{i_oJ}
$\sum_{i \neq i_o} A_i$	$p_{\cdot 1} - p_{i_o1}$	$p_{\cdot 2} - p_{i_o2}$		$p_{\cdot J} - p_{i_oJ}$

A partir de esta tabla se pueden calcular todas las medidas de asociación que se consideren apropiadas para el problema en estudio.

5.2. La Asociación entre B y Dos o Más Categorías de A.

Cuando se quiere medir la asociación entre B y el subconjunto de categorías $\{A_{i_1}, A_{i_2}, \ldots, A_{i_S}\}$ con $S \geq 2$ se puede proceder de dos maneras dependiendo de cual sea la poblacion que se quiera estudiar. Esta población puede ser la que se clasificó en la tabla de I X J o únicamente el conjunto de individuos que pertenecen a las categorías $A_{i_1}, A_{i_2}, \ldots, A_{i_S}$. En el primer caso, es decir cuando el objetivo es estudiar a toda la población, se construye una tabla de $(S + 1) \times J$ en la que cada categoría A_{i_k} le corresponde un renglón y el resto de las categorías de A se incluyen en otro renglón. Cuando únicamente se desea estudiar a los individuos clasificados en $A_{i_1}, A_{i_2}, \ldots, A_{i_S}$, se eliminan de la tabla todas aquellas categorías distintas de A_{i_k} con $k=1, \ldots, S$ obteniéndose una tabla de dimensión S x J en donde la suma de las probabilidades marginales es inferior a uno.

Para eliminar este problema las $p_{i_k j}$ se sustituyen por

$$\frac{p_{i_k j}}{\sum_{k=1}^{S} \sum_{j} p_{i_k j}}$$

6. BIBLIOGRAFIA.

1. Goodman, L. A., Kruskal, W. H.:
Measures of association for cross classifications.
II: Further discussion and references,
Journal of the American Statistical Association.
1959; 54: 123-163.

VI. MEDIDAS DE CONFIABILIDAD.

1. Medidas de confiabilidad que no excluyen el azar.

2. Medidas de confiabilidad que excluyen el azar.

3. Indices que miden el desacuerdo.

4. Bibliografía.

VI. MEDIDAS DE CONFIABILIDAD

Para determinar si una población fue clasificada correctamente respecto a una variable nominal se puede comparar la clasificación en cuestión con otra realizada de manera independiente por otro individuo (o método de asignación a clases). Gráficamente las dos clasificaciones se pueden presentar en una tabla como la siguiente.

la. Clasificación	A_1	2a. Clasificación A_1		A_i		A_J
	A_1	p_{11}		p_{1i}		p_{1J}
	A_i	p_{i1}		p_{ii}		p_{iJ}
	A_I	p_{I1}		p_{Ii}		p_{IJ}

En esta tabla el número de categorías en los renglones y en las columnas es el mismo; además, en ambos casos las categorías son iguales y están ordenadas de la misma manera. Los individuos que pertenecen a la diagonal principal de la tabla son los que fueron clasificados en la misma categoría por los dos métodos de asignación a clases. Entre mayor sea el número de individuos en la diagonal principal, mayor será la confiabilidad entre los dos métodos. Es decir, entre mayor sea el número de individuos clasificados en las celdas (A_i, A_i) con $i = 1, \ldots, I$ más confiable será la clasificación puesta a -

prueba.

Las medidas de confiabilidad también se usan cuando se quiere comparar la respuesta de un individuo que pertenece a una población con la respuesta dada por otro individuo con el que está íntimamente relacionado y que pertenece a una población distinta. Las dos poblaciones pueden ser, por ejemplo, padres e hijos. Si a estas dos poblaciones se les pregunta quién es el que generalmente inicia la conversación durante la cena: el padre, la madre o el hijo, la tabla que se obtendría sería:

B: El hijo opina que generalmente la conversación la inicia:

A: El padre opina que generalmente la conversación la inicia:

	El Padre	La Madre	El Hijo
El Padre	p_{11}	p_{12}	p_{13}
La Madre	p_{21}	p_{22}	p_{23}
El Hijo	p_{31}	p_{32}	p_{33}

Si todas las parejas de padres e hijos coinciden en su respuesta los datos se concentrarán en la diagonal principal de la tabla.

Todas las medidas de confiabilidad están basadas en $\sum_i p_{ii}$, la proporción de datos en la diagonal principal de la tabla. La diferencia más aparente entre las medidas de confiabilidad es que en algunas no se toma en cuenta a los individuos

que por azar fueron clasificados en la diagonal principal de la tabla mientras que otras sí.

1. MEDIDAS DE CONFIABILIDAD QUE NO EXCLUYEN EL AZAR.

1.1. Un Indice Meteorológico.

Una medida propuesta en 1884 para tablas de 3 x 3 - que permite determinar qué tan confiables son los - pronósticos meteorológicos es:

$$\sum_{i=1}^{3} p_{ii} + \frac{1}{2} \sum_{i \neq j=1} \sum p_{ij}$$

1.2. Indice de Klein.

El índice que definió H.J. Klein en 1885 para medir la confiabilidad en una tabla de I x I es: $\sum_i p_{ii}$; tiene el defecto de que depende de las probabilidades marginales de la tabla.

1.3. Indice de Cartwright.

Cuando una población de tamaño N es clasificada por J jueces distintos en I categorías mutuamente excluyentes, se puede usar el índice propuesto por D. S. Cartwright:

$$\frac{2}{J\,(J-1)} \sum_j \sum_{k>j} \sum_{i_j = i_k} p_{\cdots i_j \cdots i_k}$$

para medir la probabilidad de que dos jueces seleccionados al azar coincidan al clasificar a un individuo escogido aleatoriamente. (En la expresión -

anterior, $p_{\cdots i_j \cdots i_k}$ denota la probabilidad de que el j-ésimo y el k-ésimo juez clasifiquen al individuo en la i-ésima clase y $\frac{2}{J(J-1)}$ indica cuántas parejas distintas de jueces se pueden formar con J jueces). Cuando J=2 el índice de Cartwright se reduce a $\sum_i p_{ii}$.

1.4. La Lambda de Confiabilidad.

En 1945, Goodman y Kruskal definieron una medida de confiabilidad usando la lógica de reducción proporcional en el error. Este índice mide, en un sentido predictivo, la concordancia entre dos clasificaciones realizadas por dos métodos de asignación a clases; es decir, la medida propuesta por Goodman y Kruskal permite determinar si sabiendo cómo fue clasificado un individuo por alguno de los métodos de asignación a clases se puede predecir cómo fue clasificado por el otro.

Los dos métodos de asignación se designarán como A y B y se supondrá que clasifican a una población en I categorías nominales. El modelo probabilístico consiste en seleccionar al azar a un individuo de la población de interés. Después, mediante algún procedimiento aleatorio se establece si se predice como fue clasificado por A o B. Las predicciones se realizan tratando de minimizar, en un sentido fre-

cuentista, el número de errores que se cometen al usar las dos reglas siguientes:

1. No se tiene información acerca de cómo fue clasificado por el otro método.

y 2. Se sabe cómo fue clasificado por el método no seleccionado.

Al usar la regla 1, el individuo es asignado a la clase modal de A y B. Dicha clase se denota como A_M (o B_M) y es tal que $p_{M\bullet} + p_{\bullet M} \geq p_{i\bullet} + p_{\bullet i}$ $\forall$ i= 1,..., I. La probabilidad de equivocarse con 1 es: $P(1) = 1 - \frac{1}{2}(p_{M\bullet} + p_{\bullet M})$ si la probabilidad de predecir como fue clasificado por A o B es igual a 0.5. Con la regla 2 se sabe que alguno de los dos métodos asignó al individuo en A_i. Si las dos clasificaciones concuerdan entonces la probabilidad de cometer un error con la regla 2 es $P(2) = 1 - \sum_i p_{ii}$. Por lo tanto, la lambda de confiabilidad se define como (ver el Capítulo IV):

$$\lambda_r = \frac{\sum_i p_{ii} - \frac{1}{2}(p_{M\bullet} + p_{\bullet M})}{1 - \frac{1}{2}(p_{M\bullet} + p_{\bullet M})}$$

Las propiedades de este índice son:

1.4.1. λ_r toma valores en el intervalo $[-1,1]$.

1.4.2. $\lambda_r = 1$ si los dos métodos de asignación siempre coinciden al hacer la clasificación.

1.4.3. $\lambda_r = -1$ si los dos métodos nunca coinciden al clasificar a un individuo (es decir si $p_{ii} = 0 \ \forall$ i) y además $p_{M\cdot} + p_{\cdot M} = 1$.

1.4.3. Esta medida es indeterminada cuando los dos métodos de asignación clasifican a toda la población en una sola categoría.

1.4.4. La lambda de confiabilidad no toma ningún valor en particular cuando las asignaciones hechas por el método A son independientes de las asignaciones realizadas por el método B.

2. MEDIDAS DE CONFIABILIDAD QUE EXCLUYEN EL AZAR.

2.1. La Kappa sin Ponderar.

2.1.1. Definición.

En 1960, Cohen propuso una medida de confiabilidad para tablas de I x I llamada Kappa. Esta medida es igual a la diferencia entre la proporción de individuos en la diagonal principal de la tabla: $P_o = \sum_i p_{ii}$ y la proporción de individuos que por azar cayeron en dicha diagonal: $P_c = \sum_i p_{i\cdot}p_{\cdot i}$, normada, esta diferencia, por su valor más grande: $1-P_c$. Es decir, la Kappa se define como:

$$K = \frac{P_o - P_c}{1 - P_c} = \frac{\sum_i p_{ii} - \sum_i p_{i\cdot}\, p_{\cdot i}}{1 - \sum_i p_{i\cdot}p_{\cdot i}}$$

2.1.2. Interpretación.

Esta medida se interpreta como la proporción de individuos (o parejas) que fueron asignados (as) a la misma categoría sin tomar en cuenta el azar.

2.1.3. Ejemplo.

En una investigación realizada en Estados Unidos, se les preguntó a 3,845 matrimonios cuál era su candidato para ocupar la presidencia; el objetivo del estudio era determinar cuál era la proporción de conyuges que coincidían en cuanto al candidato elegido.

Los datos son (Reynold, 1977):

Tabla 2.1.3.

		B: Candidato de la Esposa		
		Nixon	Humphrey	Wallace
A: Candidato del Esposo	Nixon	1586	117	49
	Humphrey	103	1540	40
	Wallace	34	17	359

Se puede ver que la mayoría de las parejas eligieron el mismo candidato, puesto que los datos se concentran en la diagonal principal de la tabla. De hecho el 90.6% de los matrimonios estuvieron de acuerdo

con el candidato, incluyendo a los que coincidieron por azar. Las medidas de confiabilidad calculadas para la tabla 2.1.3., así como sus frecuencias marginales son:

FRECUENCIAS MARGINALES.

NIVEL	RENGLONES	COLUMNAS
1	1752.0	1723.0
2	1683.0	1674.0
3	410.0	448.0

MEDIDAS DE CONFIABILIDAD.

LA KAPPA SIN PONDERAR.	0.8421
LA LAMBDA DE GOODMAN Y KRUSKAL.	0.8292

Tanto la kappa como la lambda indican que el grado de confiabilidad entre los matrimonios es fuerte. La primera señala que 84.21% de los matrimonios coincidieron en el candidato elegido, sin tomar en cuenta a los que coincidieron por azar. La lambda de confiabilidad es menor que la kappa e indica que si se sabe qué candidato eligió uno de los conyuges; la proporción de errores que se eliminan al predecir el candidato del otro es igual a 0.8292.

A continuación se enuncian algunas propiedades de K así como su estimador de máxima -

verosimilitud y la varianza asintótica de éste.

2.1.4. Propiedades.

2.1.4.1. $K = 0$ si las variables son independientes, es decir si $p_{ii} = p_{i\cdot}p_{\cdot i} \forall i=1,\ldots,I$.

2.1.4.2. $K = 1$ si $\sum_i p_{ii} = 1$; en otras palabras, $K = 1$ si todas las parejas están de acuerdo o los dos métodos de asignación a clases siempre coinciden.

2.1.4.3. $K < 0$ si $\sum_i p_{i\cdot}p_{\cdot i} > \sum_i p_{ii}$.

2.1.5. Estimador y Varianza Asintótica.

El estimador de máxima verosimilitud de K es:

$$\hat{K} = \frac{\sum_i N_{ii} - \frac{1}{n}\sum_i N_{i\cdot}N_{\cdot i}}{n - \frac{1}{n}\sum_i N_{i\cdot}N_{\cdot i}} = \frac{n\sum_i N_{ii} - \sum_i N_{i\cdot}N_{\cdot i}}{n^2 - \sum_i N_{i\cdot}N_{\cdot i}}$$

La varianza asintótica de $\hat{K}$ es:

$$Var(\hat{K}) = \frac{1}{n}\left\{\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4-4\theta_2^2)}{(1-\theta_2)^4}\right.$$

$$\theta_1 = \sum_i p_{ii}$$

$$\theta_2 = \sum_i p_{i\cdot}p_{\cdot i}$$

$$\theta_3 = \sum_i p_{ii}(p_{i\cdot} + p_{\cdot i})$$

$$\theta_4 = \sum_i \sum_j p_{ij}(p_{j\cdot} + p_{\cdot i})^2$$

El teorema 2.1.6. señala cual es la varianza asintótica de $\hat{k}$ cuando las variables A y B son independientes.

2.1.6. Teorema.

Si las variables son independientes entonces la varianza asintótica de $\hat{k}$ es igual a:

$$Var(\hat{k}) = \frac{1}{n(1-\theta_2)^2}\left[\theta_2(1-\theta_2) - \sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i})\right]$$

Demostración:

Si $p_{ij} = p_{i\cdot}p_{\cdot j}\ \forall\ i,j$ entonces $\theta_1 = \theta_2$

por lo tanto, $Var(\hat{k}) = \frac{1}{n(1-\theta_2)^2}(\theta_2 - \theta_2^2 + 4\theta_2^2 - 2\theta_3 + \theta_4 - 4\theta_2^2)$

Además, $\theta_3 = \sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i})$ y $\theta_4 = \sum_i\sum_j p_{i\cdot}p_{\cdot j}(p_{j\cdot}+p_{\cdot i})^2$

entonces, $-2\theta_3 + \theta_4 = -2\sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i}) + \sum_i\sum_j p_{i\cdot}p_{\cdot j}(p_{j\cdot}+p_{\cdot i})^2$

$$= -2\sum_i p_{i\cdot}^2 p_{\cdot i} - 2\sum_i p_{i\cdot}p_{\cdot i}^2 + \sum_i\sum_j p_{i\cdot}p_{\cdot j}(p_{j\cdot}^2 + 2p_{j\cdot}p_{\cdot i} + p_{\cdot i}^2)$$

$$= -2\sum_i p_{i\cdot}^2 p_{\cdot i} - 2\sum_i p_{i\cdot}p_{\cdot i}^2 + \sum_i p_{i\cdot}\left(\sum_j p_{\cdot j}p_{j\cdot}^2\right) + 2\left(\sum_i p_{i\cdot}p_{\cdot i}\right)\left(\sum_j p_{\cdot j}p_{j\cdot}\right) + \left(\sum_i p_{i\cdot}p_{\cdot i}^2\right)\sum_j p_{\cdot j}$$

$$= -2\sum_i p_{i\cdot}^2 p_{\cdot i} - \sum_i p_{i\cdot}p_{\cdot i}^2 + \sum_i p_{\cdot i}^2 p_{i\cdot} + 2\left(\sum_i p_{i\cdot}p_{\cdot i}\right)^2$$

$$= -\sum_i p_{i\cdot}^2 p_{\cdot i} - \sum_i p_{i\cdot}p_{\cdot i}^2 + 2\theta_2^2 = -\sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i}) + 2\theta_2^2$$

$$\therefore\ -2\theta_3 + \theta_4 = -\sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i}) + 2\theta_2^2$$

Por lo tanto, $Var(\hat{k}) = \frac{1}{n(1-\theta_2)^2}\left[\theta_2 - \theta_2^2 + 2\theta_2^2 - \sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i})\right]$

$$\Rightarrow\ Var(\hat{k}) = \frac{1}{n(1-\theta_2)^2}\left[\theta_2(1+\theta_2) - \sum_i p_{i\cdot}p_{\cdot i}(p_{i\cdot}+p_{\cdot i})\right]$$

□

2.2. La Kappa Ponderada.

2.2.1. Introducción.

Se dice que hay un desacuerdo cuando dos métodos de asignación a clases no clasifican de la misma manera a un individuo o cuando los integrantes de una pareja tienen actitudes distintas. Teóricamente algunos desacuerdos pueden ser más graves que otros. Por ejemplo, supóngase que dos jueces clasifican de manera independiente a un grupo de personas como:

T: Sufre de algún trastorno de la personalidad.

N: Neurótico o

P: Psicótico.

El desacuerdo que existe cuando un individuo es clasificado como neurótico por un juez y como psicótico por otro juez puede ser más grave que cuando uno afirma que es neurótico y el otro que sufre de algún trastorno de la personalidad. Por esto, a veces es conveniente asignarle a cada celda de la tabla un peso que refleje el grado de desacuerdo. Estos pesos deben estar dados en una escala de razón con el fin de que sean interpreta--

bles. En la tabla 2.2.1. (Cohen, 1968) el - peso de cada celda se señala entre paréntesis.

Tabla 2.2.1.

		JUEZ A			
		T	N	P	Totales
JUEZ B	T	(0) .44	(1) .07	(3) .09	.60
	N	(1) .05	(0) .20	(6) .05	.30
	P	(3) .01	(6) .03	(0) .06	.10
Totales		.50	.30	.20	1

El peso de las celdas en la diagonal principal de la tabla anterior es cero porque en - dichas celdas no hay desacuerdo entre los - jueces. En cambio, el desacuerdo en las celdas marcadas con 6 es dos veces mayor que en aquellas marcadas con 3.

Cabe señalar que en algunos problemas puede ser conveniente ponderar la concordancia y no el desacuerdo. De cualquier manera, el índice propuesto por Cohen para medir la proporción ponderada de individuos concordantes, sin tomar en cuenta el azar, es:

2.2.2. Definición.

La kappa ponderada se define como:

$$k_\omega = \frac{P_o' - P_c'}{1 - P_c'}$$

en donde: P_o' es una proporción ponderada - que indica cuál es la confiabilidad observada en una tabla - de I x I

y P_c' es una proporción ponderada - que mide la concordancia debida al azar.

2.2.3. Ejemplo.

Si, en la tabla 2.2.1., v_{ij} representa el grado de desacuerdo en la celda (A_i, B_j) con $i,j = 1,2,3$ entonces:

$$P_o' = 1 - \sum_i \sum_j v_{ij}\, p_{ij} = 1 - 0(.44) - 1(.07) - \ldots - 6(.03) - 0(.06) = 1 - 0.9 = 0.1$$

y $$P_c' = 1 - \sum_i \sum_j v_{ij}\, p_{i\cdot} p_{\cdot j} = 1 - 0(.5).6 - 1(.3).6 - \ldots - 0(.2).1 = 1 - 1.38 = -0.38$$

Por lo tanto:

$$K_\omega = \frac{0.1 + 0.38}{1 + 0.38} = \frac{0.48}{1.38} \quad 0.348$$

2.3. La Confiabilidad Condicional.

2.3.1. Introducción.

La kappa sin ponderar y la kappa ponderada -

son dos medidas globales de la confiabilidad que existe en una tabla de I x I; no permiten determinar entre qué parejas de individuos existe mayor confiabilidad o cuándo dos métodos de asignación a clases coinciden con mayor frecuencia. Este tipo de asociación se denomina confiabilidad condicional y, en la tabla 2.1.3., permite saber cuál es la concordancia que existe entre aquellos matrimonios en los que uno de los cónyuges votó por Nixon, Wallace o Humphrey.

2.3.2. Definición y Estimador.

El índice propuesto por Colenan y Light para medir la confiabilidad entre A y B considerando, únicamente a los individuos que fueron clasificados en A_i es:

$$K_i = \frac{\left(\frac{p_{ii}}{p_{i\cdot}} - p_{\cdot i}\right)}{1 - p_{\cdot i}} = \frac{p_{ii} - p_{\cdot i}\, p_{i\cdot}}{p_{i\cdot} - p_{i\cdot}\, p_{\cdot i}}$$

Si al numerador y al denominador de K_i se les agrega la suma sobre "i" se obtiene la kappa sin ponderar.

El estimador de máxima verosimilitud de K_i es:

$$\hat{K}_i = \frac{nN_{ii} - N_{i\cdot}N_{\cdot i}}{nN_{i\cdot} - N_{i\cdot}N_{\cdot i}}$$

A continuación se explica, mediante un ejemplo, cómo se interpreta $\hat{K}_i$.

2.3.3. Ejemplo.

En la siguiente tabla de contingencia (Bishop, 1975) se muestra cómo clasificaron 2 supervisores a 72 profesores. Se considerará que el supervisor 1 tiene más experiencia que el

2 y que el objetivo es determinar cuándo existe mayor concordancia entre ellos.

Tabla 2.3.3.

B: Supervisor 2

A: Supervisor 1	Autoritario	Democrático	Permisivo	Totales
Autoritario	17	4	8	29
Democrático	5	12	0	17
Permisivo	10	3	13	26
Totales	32	19	21	72

Al calcular la medida de Coleman y Light para el subconjunto de profesores que el supervisor 1 clasificó como autoritarios se tiene que:

$$\hat{k}_1 = \frac{72(17) - 29(32)}{72(29) - 29(32)} = \frac{296}{1160} = 0.2551$$

Esto significa que sin tomar en cuenta el azar, solo 25.51% de los 29 profesores que el supervisor 1 consideró como autoritarios fueron clasificados como tales por el supervisor 2. De manera análoga se interpretan $\hat{k}_2 = 0.6004$ y $\hat{k}_3 = 0.2941$. Como $\hat{k}_2$ es el índice más grande entonces el supervisor 2 -

2 y que el objetivo es determinar cuándo existe mayor concordancia entre ellos.

Tabla 2.3.3.

A: Supervisor 1 \ B: Supervisor 2	Autoritario	Democrático	Permisivo	Totales
Autoritario	17	4	8	29
Democrático	5	12	0	17
Permisivo	10	3	13	26
Totales	32	19	21	72

Al calcular la medida de Coleman y Light para el subconjunto de profesores que el supervisor 1 clasificó como autoritarios se tiene que:

$$\hat{k}_1 = \frac{72(17) - 29(32)}{72(29) - 29(32)} = \frac{296}{1160} = 0.2551$$

Esto significa que sin tomar en cuenta el azar, solo 25.51% de los 29 profesores que el supervisor 1 consideró como autoritarios fueron clasificados como tales por el supervisor 2. De manera análoga se interpretan $\hat{k}_2 = 0.6004$ y $\hat{k}_3 = 0.2941$. Como $\hat{k}_2$ es el índice más grande entonces el supervisor 2

está más de acuerdo con el supervisor 1 al -- clasificar a los profesores como democráticos

2.3.4. Varianza Asintótica.

La varianza asintótica de $\hat{K}_i$ es:

$$\text{Var}(\hat{k}_i) = \frac{(p_{i\cdot} - p_{ii})}{np_{i\cdot}^3 (1-p_{\cdot i})^3}\left[(p_{i\cdot}-p_{ii})(p_{i\cdot}p_{\cdot i}-p_{ii})+p_{ii}(1-p_{i\cdot}-p_{\cdot i}+p_{ii})\right]$$

Esta expresión se simplifica cuando A_i es independiente de B (teorema 2.3.5.).

2.3.5. Teorema.

Si la i-ésima categoría de A es independiente de B entonces: $\text{Var}(\hat{K}_i) = \dfrac{p_{\cdot i}(1-p_{i\cdot})}{np_{i\cdot}(1-p_{\cdot i})}$

Demostración.

Como por hipótesis $p_{ii} = p_{i\cdot}p_{\cdot i}$ entonces,

$$\text{Var}(\hat{K}_i) = \frac{(p_{i\cdot} - p_{i\cdot}p_{\cdot i})}{np_{i\cdot}^3(1-p_{\cdot i})^3}\left[p_{i\cdot}p_{\cdot i}(1-p_{i\cdot}-p_{\cdot i}+p_{i\cdot}p_{\cdot i})\right]$$

$$= \frac{p_{i\cdot}^2(1-p_{\cdot i})p_{\cdot i}}{np_{i\cdot}^3(1-p_{\cdot i})^3}\left[1-p_{i\cdot}(1-p_{\cdot i})-p_{\cdot i}\right]$$

$$= \frac{p_{\cdot i}}{np_{i\cdot}(1-p_{\cdot i})^2}(1-p_{i\cdot})(1-p_{\cdot i})$$

$$\Rightarrow \text{Var}(\hat{K}_i) = \frac{p_{\cdot i}(1-p_{i\cdot})}{np_{i\cdot}(1-p_{\cdot i})}$$

□

3. INDICES QUE MIDEN EL DESACUERDO.

En algunos casos es importante medir el desacuerdo entre dos métodos de asignación a clases o entre las parejas en estudio puesto que se sabe que la confiabilidad entre ellos es fuerte. Por ejemplo, en la siguiente tabla de contingencia (Bishop, 1975) se puede ver que, en la mayoría de los casos, los reportes dados por las personas que fueron víctimas de algún crimen coinciden con el reporte dado por el policía que las socorrió.

A: Policía \ B: Víctima	Asalto	Allanamiento Morada	Latrocinio	Robo	Estupro
Asalto	33	0	0	5	1
Allanamiento Morada	0	91	2	0	0
Latrocinio	0	12	56	0	0
Robo	0	0	6	54	0
Estupro	5	0	0	0	25

En esta tabla, la discrepancia más fuerte se presenta cuando la víctima dijo haber sido asaltada; 38 asaltos fueron reportados por las víctimas; de éstos, 13.15% fueron considerados como estupro por los policías.

El desacuerdo se puede medir condicionando λ_i, λ_j y λ a que $p_{ii} = 0 \; \forall \; i = 1,\ldots,I$. (Esto es equivalente a sustituir todas las p_{ii} por cero y las p_{ij} con $i \neq j$ por

$\frac{p_{ij}}{1 - \sum_i p_{ii}}$. Otro índice que se puede usar para medir - el desacuerdo es: $\sum_i \sum_j \frac{\left(\frac{N_{ij}}{n} - \hat{p}_{ij}\right)^2}{\hat{p}_{ij}}$ en donde $\hat{p}_{ij}$ es el - estimador de máxima verosimilitud de

$$p_{ij} = \begin{cases} p_{ii} & \text{si } i=j \\ a_i b_j & \text{si } i \neq j \end{cases}$$

Esta última medida se parece a la estadística ji-cuadrada.

4. BIBLIOGRAFIA.

1. Bishop, Y. M., Fienberg, S., Hollend, P.W.:
Discrete Multivariate Analysis : Theory and Practice.
MIT Press, Cambridge Mass., 1975.

2. Cohen, J.:
Wighted Kappa
Psychological Bulletin, 1968; 70: 213-220.

3. Conger, A.J.:
Integration and generalization of kappas for multiple raters.
Psychological Bulletin, 1980; 88: 322-328

4. Fleiss, J. L., Cohen, J., Everitt, B. S.:
Large sample standard errors of kappa and weighted kappa.
Psychological Bulletin, 1969; 72: 323-327.

5. Fleiss, J. L.:
Measuring nominal scale agreement among may raters.
Psychological Bulletin, 1971; 76: 378-382.
Measuring agreement between two judges on the presence and abscence of a trait.
Biometrics, 1975; 31: 651-659.

6. Fleiss, J. L.:
Statistical Methods for Rates and Proportions
John Wiley, New York, 1981.

7. Goodman, L. A., Kruskal, W. H.:
Measures of association for cross classifications.
Journal of the American Statistical Association,
1954; 49: 732-763.
Measures of association for cross classifications.
II: Further Discussion and References.
Journal of the American Statistical Association,
1959; 54: 123-163.

8. Light, R.J.:
Measures of response agreement for qualitative data:
some generalizations and alternatives.
Psychological Bulletin, 1971; 76: 365-377.

9. Reynolds, H. T.:
The Analysis of Cross-Classifications.
The Free Press, New York, 1977.

APENDICE

LOS PROGRAMAS DE COMPUTO: RVENTAJA Y DATCUALIT

APENDICE

LOS PROGRAMAS DE COMPUTO: RVENTAJA Y DATCUALIT

1. Introducción.

Los programas de cómputo: RVENTAJA y DATCUALIT calculan algunas medidas de asociación para dos variables nominales. Dos de estos índices miden la confiabilidad, los otros están basados en la estadística ji-cuadrada, el cociente de productos cruzados y la lógica de reducción proporcional en el error.

RVENTAJA Y DATCUALIT están escritos en Pascal Estándar y fueron creados en la APPLE-II. Para usarlos es necesario tener dos discos; uno se llama # 4 y el otro # 5. El # 4 siempre se coloca en el "drive" 1 de la APPLE-II y contiene el código en binario de los dos programas así como el sistema operativo de la APPLE-II. El disco # 5 se coloca en el "drive" 2 y en él se graban los archivos de resultados y el de datos.

2. Cómo Correr los Programas.

Para poner en funcionamiento a RVENTAJA y DATCUALIT los discos deben estar colocados en el "drive" que les corresponde. Después se enciende la máquina y aparece en la parte superior de la pantalla:

COMAND: E(DIT, R(UN, F(ILE, L(INK, X(ECUTE, A(SSEM, D(EBUG,?

Estas son algunas de las funciones que puede realizar el sistema operativo. Para que se ejecute (o corra) un -

programa hay que oprimir la tecla marcada como X. En seguida hay que teclear el nombre del código en binario del programa a usar. Por ejemplo, para utilizar RVENTAJA hay que teclear # 4: ARCH1.CODE y para manejar DATCUALIT hay que escribir # 4: ARCH2.CODE.

3. El Archivo de Datos y los de Resultados.

Mediante el programa RVENTAJA se pueden crear dos archivos en el # 5; uno se llama # 5: FREC.DATA.TEXT y el otro # 5: RESULT1.TEXT. El primero es un archivo de datos que contiene la información siguiente:

1. El número de renglones y columnas de la tabla de contingencia.
2. El tamaño de la muestra.

y 3. La tabla de contingencia.

En el archivo # 5: RESULT1.TEXT se graban los resultados que se obtienen al calcular las medidas basadas en la relación de ventaja.

El programa DATCUALIT solo crea un archivo de resultados llamado # 5: RESULT2.TEXT que contiene el valor de las frecuencias marginales de la tabla así como las medidas basadas en la estadística ji-cuadrada, la lógica de reducción proporcional en el error y las de confiabilidad.

4. Cómo Imprimir un Archivo.

Para imprimir los archivos de datos o de resultados se teclea F a nivel de comandos. (Se está en el nivel de -

comandos cuando en la pantalla se observa la línea:

COMAND: E(DIT, R(UN, F(ILE, L(INK, X(ECUTE, A(SSEM, D(EBUG,?).

Al teclear la F aparece en la parte superior de la pantalla:

FILER: G,S,N,L,R,C,T,D,Q.

La opción T sirve para mandar a la impresora o a la pantalla un archivo tipo TEXT. Al oprimir la tecla marcada como T el sistema operativo pregunta ¿qué archivo voy a mandar? (TRANSFER WHTA FILE?); para ello debe escribirse el nombre del archivo que se quiere mandar a la impresora o a la pantalla es decir # 5: FREC.DATA.TEXT o # 5: RESULT1.TEXT o # 5: RESULT2.TEXT. En seguida, el sistema operativo pregunta a dónde hay que mandar el archivo. Si se contesta # 6: o PRINTER el archivo es mandado a la impresora y si se contesta CONSOLE el archivo aparece en la pantalla. (Para regresar al nivel de comandos se teclea Q).

5. Restricciones de los Programas.

En la APPLE-II los archivos de datos solo pueden leerse de manera secuencial; es decir, no se puede obtener ningún dato de manera directa, hay que recorrer todo el archivo hasta llegar al dato que se quiere usar. Por esta razón los programas fueron hechos para calcular algunas medidas de asociación entre, únicamente, dos variables.

El número de categorías que puede tener cada variable está especificado por la constante NMAXNIV que es igual a diez. Para trabajar con tablas de contingencia que tengan más de diez renglones o columnas hay que modificar el valor de NMAXNIV.

6. Variables Globales de RVENTAJA y DATCUALIT.

RVENTAJA Y DATCUALIT tienen en común las siguientes variables globales:

ARCHENT y ARCHSAL: la primera se refiere al archivo de datos y la segunda al de resultados.

FREC: esta variable guarda en memoria el valor de una sola frecuencia de la tabla de contingencia. Esta frecuencia debe ser siempre un número entero menor que 32767.

I: es un índice que sirve para pasar de un renglón a otro de la tabla.

J: este índice se usa para cambiar de una columna a otra de la tabla.

K: esta variable solo toma dos valores: cero y uno; el usuario del programa determina cuál de estos dos valores toma K. Si el programa debe ejecutar alguna tarea, el valor de K debe ser uno; de lo contrario K debe ser igual a cero.

NUMNIV: este es un arreglo (o un vector) con dos entradas; la primera sirve para guardar en me-

moria el número de renglones que tiene la tabla de contingencia, la segunda guarda en memoria el número de columnas.

Otras variables globales de DATCUALIT son:

MENOR: guarda en memoria el valor de la entrada más chica de NUMNIV. Es decir, en MENOR se guarda el número de categorías que tiene la variable con el menor número de clases.

MAYOR: guarda en memoria el valor de la entrada más grande de NUMNIV. En otras palabras, MAYOR indica cuántas categorías tiene la variable que cuenta con el mayor número de clases. Si las dos variables nominales tienen el mismo número de categorías se le asigna a MENOR el número de renglones que tiene la tabla de contingencia.

M: ésta es una variable entera que guarda en memoria el tamaño de la muestra.

MARG: es un arreglo de dimensión 2XNMAXNIV que guarda el valor de las frecuencias marginales de la tabla.

En las secciones 7 y 8 se explica como funcionan -

RVENTAJA y DATCUALIT.

7. RVENTAJA.

El programa RVENTAJA crea el archivo de datos # 5: - FREC.DATA.TEXT y calcula algunas medidas de asociación - basadas en la relación de ventaja; sus tres procedimientos principales son: CARATULA, DAMEDATOS y COCIENTE. De - estos tres procedimientos, el primero que se ejecuta es - CARATULA, que escribe en la pantalla la información siguiente:

```
.....................................................
PROGRAMO:    REBECA AGUIRRE HERNANDEZ.
LUGAR:       FACULTAD DE CIENCIAS,UNAM.
TEMA:        MEDIDAS DE ASOCIACION NOMINALES.
FECHA:       JULIO DE 1986.
.....................................................
```

```
                    PARTE 1
          CREA UN ARCHIVO DE DATOS  Y
          CALCULA ALGUNOS INDICES BASADOS EN
          EL COCIENTE  DE PRODUCTOS CRUZADOS
```

Luego pregunta: VAS A METER LOS DATOS? (SI=1, NO=0).

Cuando no se quiere crear un archivo de datos debe - teclearse el número cero y el retorno de línea; al teclear el número uno y el retorno de línea se ejecuta el procedimiento DAMEDATOS. Este procedimiento pide los datos en - la pantalla, los lee y los guarda en el archivo # 5: - FREC.DATA.TEXT. Para este fin, DAMEDATOS está integrado - por dos procedimientos: DATGRALS y FRECUEN.DATGRALS es

el primero que se ejecuta y empieza anunciándole al usuario cuál es el número máximo de categorías que pueden tener las variables en estudio. Luego pregunta: CUANTOS NIVELES TIENE LA VARIABLE 1? a esto se contesta con el número de renglones que tiene la tabla de contingencia. Después, DATGRALS pregunta por el número de niveles de la variable 2 y la respuesta debe ser el número de columnas de la tabla de contingencia. El último dato que se pide en este procedimiento es el tamaño de la muestra. Posteriormente FRECUEN pide las frecuencias de la tabla de contingencia. Estas frecuencias se deben teclear renglón por renglón cuidando de marcar un retorno de línea después de cada dato. Es decir, primero se teclea la frecuencia de la celda (1,1) luego se marca un retorno de línea y se teclea la frecuencia de la celda (1,2), etc. La creación del archivo de datos así como la ejecución de FRECUEN y DAMEDATOS concluyen al terminar de teclear la última frecuencia de la tabla con su respectivo retorno de línea. A continuación RVENTAJA pregunta si se van a calcular las medidas de asociación basadas en la relación de ventaja. Si la respuesta es negativa RVENTAJA finaliza su ejecución; de lo contrario, CARATULA guarda en el # 5: RESULT1.TEXT la información siguiente:

```
.....................................................
PROGRAMO:    REBECA AGUIRRE HERNANDEZ.
LUGAR:       FACULTAD DE CIENCIAS,UNAM.
TEMA:        MEDIDAS DE ASOCIACION NOMINALES.
FECHA:       JULIO DE 1986.
.....................................................
```

y se lleva a cabo el procedimiento COCIENTE, que tiene tres subrutinas (o procedimientos): PRINCIPIO, CALCULO y ESCRMED2. Usando el archivo de datos # 5: FREC.DATA.TEXT, PRINCIPIO guarda en el arreglo COLUMN las frecuencias de las celdas (i,J), para i=1,...,I-1; además, guarda en el arreglo RENG las frecuencias del último renglón de la tabla. Posteriormente, CALCULO lee del archivo de datos la frecuencia de la celda (i,j) con i=1,...,I-1 y j=1,..., J-1; y verifica si el denominador del cociente de productos cruzados, α_{ij}, es distinto de cero. En caso de que el denominador de α_{ij} sea cero se graba un letrero en el archivo de resultados que dice: "α_{ij} es indeterminada". Si el denominador de α_{ij} es distinto de cero la función COCIEN obtiene el valor de α_{ij}. Además, cuando $\alpha_{ij} \neq 0$ la función LOGNAT calcula el logaritmo natural de α_{ij} y si la tabla es de 2 x 2 se obtiene, mediante las funciones QYULE y YYULE, el valor de la Q de Yule y la Y de Yule. Finalmente se ejecuta el procedimiento ESCRMED2; este procedimiento está integrado por dos subrutinas: AESCRIBO y BESCRIBO que graban los resultados en el archivo # 5: RESULT1.TEXT. La subrutina AESCRIBO se emplea cuando la tabla es de 2 x 2; en otro caso se ejecuta BESCRIBO. De esta manera concluye COCIENTE; después se cierran los archivos de datos y de resultados y aparece en la pantalla el letrero:

FIN DE LA PARTE 1

Esto le indica al usuario que el programa RVENTAJA terminó de ejecutarse.

8. DATCUALIT.

El programa DATCUALIT calcula las frecuencias marginales de la tabla de contingencia grabada en el archivo de datos. Además, calcula dos medidas de confiabilidad (la kappa sin ponderar y la lambda de confiabilidad) así como algunas medidas de asociación basadas en la estadística ji-cuadrada y la lógica de reducción proporcional en el error. Lo primero que hace este programa es abrir el archivo de datos y reservar espacio en el # 5 para guardar los resultados. Luego aparece en la pantalla el anuncio:

MEDIDAS DE ASOCIACION NOMINALES

PARTE 2

y se ejecutan los procedimientos LEE, MAXMIN, MARGIN e INDICES. El primero lee del archivo de datos el número de renglones y columnas que tiene la tabla de contingencia así como el tamaño de la muestra. MAXMIN guarda en MENOR la entrada más chica del vector NUMNIV y en MAYOR la entrada más grande. Si, numéricamente, las dos entradas de NUMNIV son iguales su valor es asignado a MENOR. La sub-

rutina MARGIN calcula las frecuencias marginales de la tabla de contingencia y luego las graba en el archivo de resultados: # 5: RESULT2.TEXT. MARGIN está formada por tres procedimientos: INICIALIZA, CALCMARGIN y ESCRIBE. El primero inicia, en ceros, el arreglo MARG. CALCMARGIN calcula las frecuencias marginales y ESCRIBE las graba en el archivo de resultados. Seguidamente, la subrutina INDICES escribe en la pantalla:

ESTE PROGRAMA CALCULA ALGUNOS INDICES PERTENECIENTES A LOS SIGUIENTES GRUPOS:

A. MEDIDAS BASADAS EN LA ESTADISTICA JI-CUADRADA PARA PROBAR INDEPENDENCIA.

B. MEDIDAS BASADAS EN LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR.

C. MEDIDAS DE CONFIABILIDAD (TABLAS CUADRADAS).

y después pregunta:

A QUE GRUPO PERTENECE LA MEDIDA QUE DEBO CALCULAR ?

1. Al teclear "A" y el retorno de línea se ejecuta el procedimiento JICUAD. Su función es calcular la estadística ji-cuadrada y las medidas de asociación basadas en ella siempre y cuando todas las frecuencias marginales sean distintas de cero. (Si alguna frecuencia marginal es igual a cero el procedimiento JICUAD únicamente escribe en el archivo de resultados:

MEDIDAS BASADAS EN LA JI-CUADRADA.

ESTADISTICA JI-CUADRADA: INDEFINIDO.).

2. Si se teclea "B" y el retorno de línea se lleva a cabo el procedimiento RPE que calcula algunos índices basados en la lógica de reducción proporcional en el error. Lo primero que hace RPE es escribir en la pantalla:

QUE INDICE DEBO CALCULAR ?

LA LAMBDA DE GOODMAN Y KRUSKAL.

A. PREDICCION DE LAS COLUMNAS.

B. PREDICCION DE LOS RENGLONES.

C. PREDICCION DE COLUMNAS O RENGLONES

Mediante la opción A se obtiene el valor de $\hat{\lambda}_j$, mediante B el de $\hat{\lambda}_i$ y con C el de $\hat{\lambda}$. El resultado lo graba la subrutina ESCRMED3 en el archivo # 5: RESULT2. TEXT. Posteriormente, RPE pregunta:

QUIERES QUE CALCULE OTRO INDICE DE RPE ? (SI=1/NO=0)

Al teclear el uno (y el retorno de línea) se pregunta:

CUAL ? A, B o C.

(es decir, $\hat{\lambda}_i$, $\hat{\lambda}_j$ o $\hat{\lambda}$) y al teclear el cero el procedimiento RPE llega a su fin.

3. Por último, al teclear la opción C se verifica si la tabla de contingencia es cuadrada; en caso de que así sea la subrutina CONFIAB calcula la kappa sin ponderar y la lambda de confiabilidad. Si la tabla de contin-

gencia no es cuadrada aparece en la pantalla un letrero que dice:

LA TABLA NO ES CUADRADA.

Cuando los procedimientos JICUAD, RPE y CONFIAB concluyen su ejecución la subrutina INDICES pregunta:

QUIERES QUE CALCULE OTRO INDICE.

A. BASADO EN LA ESTADISTICA JI-CUADRADA PARA PROBAR - INDEPENDENCIA.

B. BASADO EN LA LOGICA DE REDUCCION PROPORCIONAL EN - EL ERROR.

C. DE CONFIABILIDAD ?

SI=1 ; NO=0

Si la respuesta es afirmativa INDICES pregunta:

A QUE GRUPO PERTENECE LA MEDIDA QUE DEBO CALCULAR ?

y si la respuesta es negativa, INDICES llega a su fin. En este momento DATCUALIT cierra los archivos de datos y resultados; luego aparece en la pantalla el mensaje:

FIN DE LA PARTE 2

Se anexa una copia de los programas RVENTAJA y DATCUALIT.

```
(*$L "CONSOLE:" *)
PROGRAM RVENTAJA(INPUT,OUTPUT);
USES TRANSCEND;
                    (* PROGRAMO:   REBECA AGUIRRE HERNANDEZ.
                       LUGAR:      FACULTAD DE CIENCIAS, UNAM.
                       FECHA:      JULIO DE 1986.
                       TEMA:       MEDIDAS DE ASOCIACION PARA VARIABLES NOMINALES (TESIS).
                       DATOS:      1. NUMERO DE CATEGORIAS POR RENGLON Y POR COLUMNA.
                                   2. TAMANIO DE LA MUESTRA.
                                   3. TABLA DE CONTINGENCIA.
                       RESULTADOS: ALGUNOS INDICES BASADOS EN:
                                   EL COCIENTE DE PRODUCTOS CRUZADOS.
                       METODO:     ESTE PROGRAMA ESTA FORMADO POR TRES SUBRUTINAS.
                                   LA PRIMERA ESCRIBE LA CARATULA EN EL ARCHIVO DE
                                   SALIDA. LA SEGUNDA CREA UN ARCHIVO DE DATOS Y EN
                                   LA TERCERA SUBRUTINA SE USAN FUNCIONES PARA CAL-
                                   CULAR ALGUNOS INDICES BASADOS EN EL COCIENTE DE
                                   PRODUCTOS CRUZADOS PARA TABLAS DE 2X2,LOS CUALES
                                   SON ESCRITOS EN EL ARCHIVO DE SALIDA. *)
  CONST
    NMAXNIV=10;
    BL=' ';
    MI=18;
  TYPE
    TNUMNIV=ARRAY[0..1] OF INTEGER;
  VAR
    ARCHENT,
    ARCHSAL:TEXT;
    VUELTA,
    FREC,
    I,
    J,
    K:INTEGER;
    NUMNIV:TNUMNIV;
  PROCEDURE CARATULA(VUELTA:INTEGER);
    CONST
      LINEA='..........';
    VAR
      CADENA1:STRING[10];
      CADENA2:STRING[32];
      LCADENA1:INTEGER;
    BEGIN (* CARATULA *)
      IF VUELTA=2
        THEN
          WRITELN(ARCHSAL,BL:MI,LINEA,LINEA,LINEA,LINEA)
        ELSE
          BEGIN (* ELSE *)
```

```
                WRITELN(BL:120);
                WRITELN(BL:8,LINEA,LINEA,LINEA,LINEA);
              END; (* ELSE *)
        FOR I:=0 TO 3 DO
            BEGIN (* FOR I *)
              CASE I OF
                    0:BEGIN (* 0 *)
                        CADENA1:='PROGRAMO: ';
                        CADENA2:='REBECA AGUIRRE HERNANDEZ.';
                      END; (* 0 *)
                    1:BEGIN (* 1 *)
                        CADENA1:='LUGAR: ';
                        CADENA2:='FACULTAD DE CIENCIAS,UNAM.';
                      END; (* 1 *)
                    2:BEGIN (* 2 *)
                        CADENA1:='TEMA: ';
                        CADENA2:='MEDIDAS DE ASOCIACION NOMINALES.';
                      END; (* 2 *)
                    3:BEGIN (* 3 *)
                        CADENA1:='FECHA: ';
                        CADENA2:='JULIO DE 1986.';
                      END; (* 3 *)
                END; (* CASE I *)
              LCADENA1:=12-LENGTH(CADENA1);
              IF VUELTA=2
                 THEN
                    WRITELN(ARCHSAL,BL:MI,CADENA1,BL:LCADENA1,CADENA2)
                 ELSE
                    WRITELN(BL:8,CADENA1,BL:LCADENA1,CADENA2);
            END; (* FOR I *)
        IF VUELTA=2
           THEN
              BEGIN (* THEN *)
                 WRITELN(ARCHSAL,BL:MI,LINEA,LINEA,LINEA,LINEA);
                 WRITELN(ARCHSAL,BL:528);
              END (* THEN *)
           ELSE
              BEGIN (* ELSE *)
                 WRITELN(BL:8,LINEA,LINEA,LINEA,LINEA);
                 WRITELN(BL:120);
              END; (* ELSE *)
    END; (* CARATULA *)
PROCEDURE DAMEDATOS(VAR NUMNIV:TNUMNIV);
    PROCEDURE DATGRALS(VAR NUMNIV:TNUMNIV);
       VAR
          M:REAL;
       BEGIN (* DATGRALS *)
          WRITELN('CADA VARIABLE PUEDE TENER A LO MAS ',NMAXNIV,' NIVELES DE CLASIFICACION.');
          WRITELN;
```

```
            FOR I:=0 TO 1 DO
                BEGIN (* FOR I *)
                    REPEAT
                        WRITELN('CUANTOS NIVELES TIENE LA VARIABLE ',I+1,' ?');
                        READLN(NUMNIV[I])
                    UNTIL (NUMNIV[I]>=1) AND (NUMNIV[I]<=NMAXNIV);
                    WRITE(ARCHENT,NUMNIV[I]:4);
                END; (* FOR I *)
            WRITELN('CUAL ES EL TAMANIO DE LA MUESTRA ?');
            READLN(M);
            WRITELN(ARCHENT,M:10:1);
        END; (* DATSGRALS *)
    PROCEDURE FRECUEN(NUMNIV:TNUMNIV);
        BEGIN (* FRECUEN *)
            WRITELN(BL:17,'DAME LAS FRECUENCIAS POR FAVOR.');
            WRITELN(BL:10,'SOLO TEN CUIDADO DE QUE TODAS SEAN MENORES O');
            WRITELN(BL:10,'IGUALES A ',MAXINT,' DE LO CONTRARIO EL PROGRAMA');
            WRITELN(BL:15,'DEJA DE EJECUTARSE MARCANDO ERROR.');
            FOR I:=0 TO NUMNIV[0] - 1 DO
                FOR J:=0 TO NUMNIV[I]-1 DO
                    BEGIN (* FOR J *)
                        READLN(FREC);
                        IF J=NUMNIV[I]-1
                            THEN
                                WRITELN(ARCHENT,FREC:6)
                            ELSE
                                WRITE(ARCHENT,FREC:6);
                    END; (* FOR J *)
            WRITELN('GRACIAS.');
        END; (* FRECUEN *)
    BEGIN (* DAMEDATOS *)
        DATGRALS(NUMNIV);
        FRECUEN(NUMNIV);
    END; (* DAMEDATOS *)
PROCEDURE COCIENTE(NUMNIV:TNUMNIV);
    TYPE
        TCOCIEN=ARRAY[2..NMAXNIV,2..NMAXNIV] OF REAL;
        TCOLUMN=ARRAY[2..NMAXNIV] OF REAL;
        TRENG=ARRAY[1..NMAXNIV] OF REAL;
        TYULE=ARRAY[0..1] OF REAL;
    VAR
        COCIEN,
        LOGNAT:TCOCIEN;
        COLUMN:TCOLUMN;
        RENG:TRENG;
        YULE:TYULE;
    FUNCTION PRODCRUZ(FREC:INTEGER;COLUMN:TCOLUMN;RENG:TRENG;NUMNIV:TNUMNIV):REAL;
        VAR
            CELDAREF:INTEGER;
```

```
    BEGIN (* PRODCRUZ *)
       CELDAREF:=NUMNIV[1];
       PRODCRUZ:=(FREC * RENG[CELDAREF])/(COLUMN[I] * RENG[J]);
    END; (* PRODCRUZ *)
FUNCTION LNAT(COCIEN:TCOCIEN):REAL;
    BEGIN (* LNAT *)
       LNAT:=LN(COCIEN[I,J+1]);
    END; (* LNAT *)
FUNCTION QYULE(COCIEN:TCOCIEN):REAL;
    BEGIN (* QYULE *)
       QYULE:=(COCIEN[I,J+1] - 1)/(COCIEN[I,J+1] + 1);
    END; (* QYULE *)
FUNCTION YYULE(COCIEN:TCOCIEN):REAL;
    VAR
       RAIZ:REAL;
    BEGIN (* YYULE *)
       RAIZ:=SQRT(COCIEN[I,J+1]);
       YYULE:=(RAIZ - 1)/(RAIZ + 1);
    END; (* YYULE *)
PROCEDURE PRINCIPIO(VAR NUMNIV:TNUMNIV;VAR COLUMN:TCOLUMN;VAR RENG:TRENG);
    BEGIN (* PRINCIPIO *)
       READ(ARCHENT,NUMNIV[0]);
       READLN(ARCHENT,NUMNIV[1]);
       FOR I:=2 TO NUMNIV[0] DO
           FOR J:=1 TO NUMNIV[1] DO
               BEGIN (* FOR J *)
                  READ(ARCHENT,FREC);
                  IF J=NUMNIV[1]
                     THEN
                        COLUMN[I]:=FREC;
               END; (* FOR J *)
       FOR J:=1 TO NUMNIV[1] DO
           BEGIN (* FOR J *)
              READ(ARCHENT,FREC);
              RENG[J]:=FREC;
           END; (* FOR J *)
    END; (* PRINCIPIO *)
PROCEDURE CALCULO(COLUMN:TCOLUMN;RENG:TRENG;NUMNIV:TNUMNIV;VAR COCIEN,LOGNAT:TCOCIEN;VAR YULE:TYULE)
    BEGIN (* CALCULO *)
       CLOSE(ARCHENT);
       RESET(ARCHENT,'#5:FREC.DATA.TEXT');
       READLN(ARCHENT);
       FOR I:=2 TO NUMNIV[0] DO
           BEGIN (* FOR I *)
              FOR J:=1 TO NUMNIV[1]-1 DO
                  BEGIN (* FOR J *)
                     READ(ARCHENT,FREC);
                     IF (COLUMN[I] <> 0) AND (RENG[J] <> 0)
                        THEN
```

```
                        BEGIN (* THEN *)
                           COCIEN[I,J+1]:=PRODCRUZ(FREC,COLUMN,RENG,NUMNIV);
                           IF COCIEN[I,J+1] <> 0
                              THEN
                                 LOGNAT[I,J+1]:=LNAT(COCIEN);
                           IF (NUMNIV[0]=2) AND (NUMNIV[1]=2)
                              THEN
                                 BEGIN (* THEN *)
                                    YULE[0]:=QYULE(COCIEN);
                                    YULE[1]:=YYULE(COCIEN);
                                 END; (* THEN *)
                        END; (* THEN *)
                  END; (* FOR J *)
               READLN(ARCHENT,FREC);
            END; (* FOR I *)
      END; (* CALCULO *)
   PROCEDURE ESCRMED2(COLUMN:TCOLUMN;RENG:TRENG;NUMNIV:TNUMNIV;COCIEN,LOGNAT:TCOCIEN;YULE:TYULE
      CONST
         ENCAB0='MEDIDAS BASADAS EN LA RELACION DE VENTAJA.';
         ENCAB1='RELACION DE VENTAJA';
         ENCAB2='LOGARITMO NATURAL DE LA RELACION DE VENTAJA';
      PROCEDURE AESCRIBO(COLUMN:TCOLUMN;RENG:TRENG;COCIEN,LOGNAT:TCOCIEN;YULE:TYULE);
         CONST
            ENCAB3='LA Q DE YULE';
            ENCAB4='LA Y DE YULE';
         BEGIN (* AESCRIBO *)
            IF (COLUMN[2] <>0) AND (RENG[1] <>0)
               THEN
                  BEGIN (* THEN *)
                     WRITELN(ARCHSAL,BL:MI,ENCAB1,BL:25,COCIEN[2,2]:6:4);
                     IF COCIEN[2,2] <> 0
                        THEN
                           WRITELN(ARCHSAL,BL:MI,ENCAB2,BL,LOGNAT[2,2]:6:4)
                        ELSE
                           WRITELN(ARCHSAL,BL:MI,ENCAB2,BL,'INDEFINIDO');
                     WRITELN(ARCHSAL,BL:MI,ENCAB3,BL:32,YULE[0]:6:4);
                     WRITELN(ARCHSAL,BL:MI,ENCAB4,BL:32,YULE[1]:6:4);
                  END (* THEN *)
               ELSE
                  WRITELN(ARCHSAL,BL:MI,ENCAB1,':',BL:2,'INDEFINIDO');
         END; (* AESCRIBO *)
      PROCEDURE BESCRIBO(COLUMN:TCOLUMN;RENG:TRENG;COCIEN,LOGNAT:TCOCIEN;NUMNIV:TNUMNIV);
         CONST
            ESP=4;
            ENCAB5='SUBINDICES';
         BEGIN (* BESCRIBO *)
            WRITELN(ARCHSAL,BL:MI,ENCAB5,BL:ESP,ENCAB1,BL:ESP,ENCAB2);
            FOR I:=2 TO NUMNIV[0] DO
               FOR J:=1 TO NUMNIV[1]-1 DO
                  BEGIN (* FOR J *)
```

```
                IF (COLUMNEI] <> 0) AND (RENGEJ] <> 0) AND (COCIENEI,J+1] <> 0)
                   THEN
                      WRITELN(ARCHSAL,BL:MI,BL:3,I-1,',',J,BL:10,COCIENEI,J+1]:10:4,BL:26,LOGNATEI,J+1]:10:
                IF (COLUMNEI]<>0) AND (RENGEJ]<>0) AND (COCIENEI,J+1]=0)
                   THEN
                      WRITELN(ARCHSAL,BL:MI,BL:3,I-1,',',J,BL:10,COCIENEI,J+1]:10:4,BL:26,'INDEFINIDO');
                IF (COLUMNEI]=0) OR (RENGEJ]=0)
                   THEN
                      WRITELN(ARCHSAL,BL:MI,BL:3,I-1,',',J,BL:10,'INDEFINIDO');
             END; (* FOR J *)
       END; (* ESCRIBO *)
    BEGIN (* ESCRMED2 *)
       WRITELN(ARCHSAL,ENCABO);
       WRITELN(ARCHSAL);
       IF (NUMNIVE0]=2) AND (NUMNIVE1]=2)
          THEN
             AESCRIBO(COLUMN,RENG,COCIEN,LOGNAT,YULE)
          ELSE
             DESCRIBO(COLUMN,RENG,COCIEN,LOGNAT,NUMNIV);
       WRITELN(ARCHSAL,BL:264);
    END; (* ESCRMED2 *)
 BEGIN (* COCIENTE *)
    PRINCIPIO(NUMNIV,COLUMN,RENG);
    CALCULO(COLUMN,RENG,NUMNIV,COCIEN,LOGNAT,YULE);
    ESCRMED2(COLUMN,RENG,NUMNIV,COCIEN,LOGNAT,YULE);
 END; (* COCIENTE *)
BEGIN (* PROGRAMA PRINCIPAL *)
   VUELTA:=1;
   CARATULA(VUELTA);
   WRITELN(BL:25,'PARTE 1');
   WRITELN;
   WRITELN(BL:17,'CREA UN ARCHIVO DE DATOS Y');
   WRITELN(BL:13,'CALCULA ALGUNOS INDICES BASADOS EN');
   WRITELN(BL:13,'EL COCIENTE DE PRODUCTOS CRUZADOS.');
   WRITELN;
   WRITELN('VAS A METER LOS DATOS ? (SI=1,NO=0).');
   READLN(K);
   WHILE K=1 DO
      BEGIN (* WHILE *)
         REWRITE(ARCHENT,'#5:FREC.DATA.TEXT');
         DAMEDATOS(NUMNIV);
         K:=0;
         CLOSE(ARCHENT,LOCK);
      END; (* WHILE *)
   WRITELN('QUIERES QUE CALCULE ALGUNOS INDICES BASADOS EN ');
   WRITELN('LA RELACION DE VENTAJA ? (SI=1,NO=0).');
   READLN(K);
   WHILE K=1 DO
      BEGIN (* WHILE *)
         RESET(ARCHENT,'#5:FREC.DATA.TEXT');
```

```
            REWRITE(ARCHSAL,'#5:RESULT1.TEXT');
            VUELTA:=2;
            CARATULA(VUELTA);
            COCIENTE(NUMNIV);
            CLOSE(ARCHENT);
            CLOSE(ARCHSAL,LOCK);
            K:=0;
          END; (* WHILE *)
      WRITELN('FIN DE LA PARTE 1.');
    END. (* PROGRAMA PRINCIPAL *)



(*$L "CONSOLE:" *)
PROGRAM DATCUALIT(INPUT,OUTPUT);
USES TRANSCEND;
                    (* PROGRAMO:   REBECA AGUIRRE HERNANDEZ.
                       LUGAR:      FACULTAD DE CIENCIAS, UNAM.
                       FECHA:      JULIO DE 1986.
                       TEMA:       MEDIDAS DE ASOCIACION PARA VARIABLES NOMINALES (TESIS).
                       DATOS:      1. NUMERO DE CATEGORIAS POR RENGLON Y POR COLUMNA.
                                   2. TAMANIO DE LA MUESTRA.
                                   3. TABLA DE CONTINGENCIA.
                       RESULTADOS: 1. FRECUENCIAS MARGINALES POR RENGLON Y POR COLUMNA.
                                   2. ALGUNOS INDICES BASADOS EN:
                                      a. LA ESTADISTICA JI-CUADRADA.
                                      b. LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR.
                                   3. DOS MEDIDAS DE CONFIABILIDAD.
                       METODO:     ESTE PROGRAMA ESTA FORMADO POR TRES PROCEDIMIENTOS
                                   PRINCIPALES. EL PRIMERO LEE LOS DATOS DE UN ARCHI-
                                   VO PREVIAMENTE CREADO, EL SEGUNDO CALCULA Y ESCRI-
                                   BE LAS FRECUENCIAS MARGINALES DE LA TABLA DE CONTIN-
                                   GENCIA Y MEDIANTE EL TERCER PROCEDIMIENTO SE CALCU-
                                   LAN Y ESCRIBEN ALGUNAS MEDIDAS DE ASOCIACION NOMINALES.
                    *)
  CONST
    NMAXNIV=10;
    BL=' ';
    MI=18;
  TYPE
    TMARG=ARRAY[1..2,1..NMAXNIV] OF REAL;
    TNUMNIV=ARRAY[0..1] OF INTEGER;
  VAR
    ARCHENT,
    ARCHSAL:TEXT;
    I,
    J,
    K,
```

```
        MENOR,
        MAYOR:INTEGER;
        FREC,
        M:REAL;
        NUMNIV:TNUMNIV;
        MARG:TMARG;
      PROCEDURE LEE(VAR M:REAL; VAR NUMNIV:TNUMNIV);
        BEGIN (* LEE *)
          FOR I:=0 TO 1 DO
            READ(ARCHENT,NUMNIV[I]);
          READLN(ARCHENT,M);
        END; (* LEE *)
      PROCEDURE MAXMIN(NUMNIV:TNUMNIV;VAR MENOR,MAYOR:INTEGER);
        BEGIN (* MAXMIN *)
          IF NUMNIV[0] <= NUMNIV[1]
            THEN
              BEGIN (* THEN *)
                MENOR:=NUMNIV[0]-1;
                MAYOR:=NUMNIV[1];
              END (* THEN *)
            ELSE
              BEGIN (* ELSE *)
                MENOR:=NUMNIV[1]-1;
                MAYOR:=NUMNIV[0];
              END; (* ELSE *)
        END; (* MAXMIN *)
      PROCEDURE MARGIN(NUMNIV:TNUMNIV; VAR MARG:TMARG;MENOR,MAYOR:INTEGER);
        PROCEDURE INICIALIZA(NUMNIV:TNUMNIV; VAR MARG:TMARG);
          BEGIN (* INICIALIZA *)
            FOR I:=0 TO 1 DO
              FOR J:=0 TO NUMNIV[I]-1 DO
                MARG[I+1,J+1]:=0;
          END; (* INICIALIZA *)
        PROCEDURE CALCMARGIN(NUMNIV:TNUMNIV;VAR MARG:TMARG);
          BEGIN (* CALCMARGIN *)
            FOR I:=0 TO NUMNIV[0]-1 DO
              BEGIN (* FOR I *)
                J:=0;
                REPEAT
                  READ(ARCHENT,FREC);
                  MARG[1,I+1]:=MARG[1,I+1] + FREC;
                  MARG[2,J+1]:=MARG[2,J+1] + FREC;
                  J:=J + 1;
                UNTIL EOLN(ARCHENT);
              END; (* FOR I *)
          END; (* CALCMARGIN *)
        PROCEDURE ESCRIBE(MARG:TMARG;NUMNIV:TNUMNIV;MENOR,MAYOR:INTEGER);
          CONST
            TIT1='FRECUENCIAS MARGINALES.';
          BEGIN (* ESCRIBE *)
```

```
            WRITELN(ARCHSAL,TIT1);
            WRITELN(ARCHSAL);
            WRITELN(ARCHSAL,BL:MI,'NIVEL','RENGLONES':13,'COLUMNAS':12);
            FOR I:=1 TO MENOR +1 DO
                WRITELN(ARCHSAL,BL:MI,I:3,BL:6,MARG[1,I]:8:1,BL:6,MARG[2,I]:8:1);
            FOR I:=MENOR + 2 TO MAYOR DO
                IF NUMNIV[0]=MENOR + 1
                   THEN
                      WRITELN(ARCHSAL,BL:MI,I:3,BL:20,MARG[2,I]:8:1)
                   ELSE
                      WRITELN(ARCHSAL,BL:MI,I:3,BL:6,MARG[1,I]:8:1);
            WRITELN(ARCHSAL,BL:264);
         END; (* ESCRIBE *)
      BEGIN (* MARGIN *)
         INICIALIZA(NUMNIV,MARG);
         CALCMARGIN(NUMNIV,MARG);
         ESCRIBE(MARG,NUMNIV,MENOR,MAYOR);
      END; (* MARGIN *)
   PROCEDURE JICUAD(NUMNIV:TNUMNIV;MARG:TMARG;M:REAL;MENOR:INTEGER);
      TYPE
         TMEDIDAS=ARRAY[0..4] OF REAL;
         TBOOLEANA=ARRAY[1..2] OF BOOLEAN;
      VAR
         MEDIDAS:TMEDIDAS;
         BOOLEANA:TBOOLEANA;
      FUNCTION ESTADJICUAD(NUMNIV:TNUMNIV;MARG:TMARG;M:REAL):REAL;
         VAR
            SUMA,
            CUADRADO:REAL;
         BEGIN (* ESTADJICUAD *)
            CLOSE(ARCHENT);
            RESET(ARCHENT,'#5:FREC.DATA.TEXT');
            READLN(ARCHENT);
            SUMA:=0;
            FOR I:=0 TO NUMNIV[0]-1 DO
                FOR J:=0 TO NUMNIV[1]-1 DO
                   BEGIN (* FOR J *)
                      READ(ARCHENT,FREC);
                      CUADRADO:=FREC * FREC;
                      SUMA:=SUMA + CUADRADO/(MARG[1,I+1] * MARG[2,J+1]);
                   END; (* FOR J *)
            ESTADJICUAD:=SUMA*M-M;
         END; (* ESTADJICUAD *)
      FUNCTION COEFCONT(M:REAL;MEDIDAS:TMEDIDAS):REAL;
         BEGIN (* COEFCONT *)
            COEFCONT:=MEDIDAS[0]/M;
         END; (* COEFCONT *)
      FUNCTION COEFKARL(MEDIDAS:TMEDIDAS):REAL;
         BEGIN (* COEFKARL *)
            COEFKARL:=SQRT(MEDIDAS[1]/(MEDIDAS[1] + 1));
```

```
    END; (* COEFKARL *)
  FUNCTION TSCHUPROV(MEDIDAS:TMEDIDAS;NUMNIV:TNUMNIV):REAL;
    VAR
      DENOMIN:REAL;
    BEGIN (* TSCHUPROV *)
      DENOMIN:=SQRT((NUMNIV[0]-1)*(NUMNIV[1]-1));
      TSCHUPROV:=SQRT(MEDIDAS[1]/DENOMIN);
    END; (* TSCHUPROV *)
  FUNCTION CRAMER(MEDIDAS:TMEDIDAS;MENOR:INTEGER):REAL;
    BEGIN (* CRAMER *)
      CRAMER:=SQRT(MEDIDAS[1]/MENOR);
    END; (* CRAMER *)
  PROCEDURE ESCRMED1(MEDIDAS:TMEDIDAS;MENOR:INTEGER);
    CONST
      TIT2='MEDIDAS BASADAS EN LA JI-CUADRADA.';
    VAR
      NOMIND:STRING[50];
      ESPACIO:INTEGER;
    BEGIN (* ESCRMED1 *)
      WRITELN(ARCHSAL,TIT2);
      WRITELN(ARCHSAL);
      FOR I:=0 TO 4 DO
        BEGIN (* FOR I *)
          CASE I OF
            0:NOMIND:='ESTADISTICA JI-CUADRADA: ';
            1:NOMIND:='COEFICIENTE DE CONTINGENCIA EN MEDIA CUADRATICA: ';
            2:NOMIND:='COEFICIENTE DE CONTINGENCIA DE KARL PEARSON: ';
            3:NOMIND:='INDICE DE TSCHUPROV: ';
            4:NOMIND:='LA V DE CRAMER: ';
          END; (* CASE I *)
          ESPACIO:=50 - LENGTH(NOMIND);
          IF ((I=4) OR (I=3)) AND (MENOR<=0)
            THEN
              WRITELN(ARCHSAL,BL:MI,NOMIND,BL:ESPACIO,BL,'INDEFINIDO')
            ELSE
              WRITELN(ARCHSAL,BL:MI,NOMIND,BL:ESPACIO,MEDIDAS[I]:11:4);
        END; (* FOR I *)
      WRITELN(ARCHSAL,BL:264);
    END; (* ESCRMED1 *)
  BEGIN (* JICUAD *)
    FOR I:=1 TO 2 DO
      BEGIN (* FOR I *)
        BOOLEANA[I]:=FALSE;
        J:=1;
        REPEAT
          IF MARG[I,J]=0
            THEN
              BOOLEANA[I]:=TRUE;
          J:=J+1
        UNTIL (BOOLEANA[I]=TRUE) OR (J > NUMNIV[I-1]);
```

```
            END; (* FOR I *)
          IF (BOOLEANA[1]=FALSE) AND (BOOLEANA[2]=FALSE)
             THEN
                BEGIN (* THEN *)
                   MEDIDAS[0]:=ESTADJICUAD(NUMNIV,MARG,M);
                   MEDIDAS[1]:=COEFCONT(M,MEDIDAS);
                   MEDIDAS[2]:=COEFKARL(MEDIDAS);
                   IF MENOR > 0
                      THEN
                         BEGIN (* THEN *)
                            MEDIDAS[3]:=TSCHUPROV(MEDIDAS,NUMNIV);
                            MEDIDAS[4]:=CRAMER(MEDIDAS,MENOR);
                         END; (* THEN *)
                   ESCRMEDI(MEDIDAS,MENOR);
                END (* THEN *)
             ELSE
                BEGIN (* ELSE *)
                   WRITELN(ARCHSAL,'MEDIDAS BASADAS EN LA JI-CUADRADA.');
                   WRITELN(ARCHSAL);
                   WRITELN(ARCHSAL,BL:MI,'ESTADISTICA JI-CUADRADA: INDEFINIDO.');
                   WRITELN(ARCHSAL,BL:264);
                END; (* ELSE *)
       END; (* JICUAD *)
    PROCEDURE RPE(NUMNIV:TNUMNIV;M:REAL;MARG:TMARG);
       CONST
          TIT3='MEDIDAS BASADAS EN LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR.';
          TO='LA LAMBDA DE GOODMAN Y KRUSKAL.';
          T1='PREDICCION DE LAS COLUMNAS.';
          T2='PREDICCION DE LOS RENGLONES.';
          T3='PREDICCION DE COLUMNAS O RENGLONES.';
       TYPE
          TMMARG=ARRAY[1..2] OF REAL;
          TMAX=ARRAY[1..NMAXNIV] OF REAL;
       VAR
          BANDERA:BOOLEAN;
          OPCION2:CHAR;
          SUMA,
          LAMB:REAL;
          CONT,
          INDICE:INTEGER;
          MMARG:TMMARG;
          MCOLUMN,
          MRENG:TMAX;
       FUNCTION MAXMARG(INDICE:INTEGER; NUMNIV:TNUMNIV;MARG:TMARG):REAL;
          VAR
             TEMPORAL:REAL;
          BEGIN (* MAXMARG *)
             TEMPORAL:=MARG[INDICE,1];
             FOR I:=2 TO NUMNIV[INDICE-1] DO
                IF MARG[INDICE,I] > TEMPORAL
```

```
                THEN
                    TEMPORAL:=MARG[INDICE,I];
        MAXMARG:=TEMPORAL;
    END; (* MAXMARG *)
PROCEDURE MAXRENG(VAR MRENG:TMAX;NUMNIV:TNUMNIV);
    PROCEDURE INICIA(VAR MRENG:TMAX;NUMNIV:TNUMNIV);
        BEGIN (* INICIA *)
            FOR I:=0 TO NUMNIV[0]-1 DO
                MRENG[I+1]:=0;
        END; (* INICIA *)
    BEGIN (* INICIA *)
        INICIA(MRENG,NUMNIV);
        CLOSE(ARCHENT);
        RESET(ARCHENT,'#5:FREC.DATA.TEXT');
        READLN(ARCHENT);
        FOR I:=0 TO NUMNIV[0]-1 DO
            FOR J:=0 TO NUMNIV[1]-1 DO
                BEGIN (* FOR J *)
                    READ(ARCHENT,FREC);
                    IF FREC > MRENG[I+1]
                        THEN
                            MRENG[I+1]:=FREC;
                END; (* FOR J *)
    END; (* INICIA *)
PROCEDURE MAXCOLUMN(VAR MCOLUMN:TMAX;NUMNIV:TNUMNIV);
    BEGIN (* MAXCOLUMN *)
        CLOSE(ARCHENT);
        RESET(ARCHENT,'#5:FREC.DATA.TEXT');
        READLN(ARCHENT);
        FOR J:=0 TO NUMNIV[1]-1 DO
            BEGIN (* FOR J *)
                READ(ARCHENT,FREC);
                MCOLUMN[J+1]:=FREC;
            END; (* FOR J *)
        FOR I:=1 TO NUMNIV[0]-1 DO
            FOR J:=0 TO NUMNIV[1]-1 DO
                BEGIN (* FOR J *)
                    READ(ARCHENT,FREC);
                    IF FREC > MCOLUMN[J+1]
                        THEN
                            MCOLUMN[J+1]:=FREC;
                END; (* FOR J *)
    END; (* MAXCOLUMN *)
FUNCTION SUM(MRENG,MCOLUMN:TMAX;NUMNIV:TNUMNIV;INDICE:INTEGER):REAL;
    VAR
        SUMO:REAL;
    BEGIN (* SUM *)
        SUMO:=0;
        IF INDICE=2
            THEN
```

```
            FOR I:=1 TO NUMNIV[0] DO
               SUMO:=SUMO + MRENG[I]
         ELSE
            FOR J:=1 TO NUMNIV[1] DO
               SUMO:=SUMO + MCOLUMN[J];
      SUM:=SUMO;
   END; (* SUM *)
FUNCTION LAMBDA(SUMA,M:REAL;INDICE:INTEGER;MMARG:TMMARG):REAL;
   BEGIN (* LAMBDA *)
      LAMBDA:=(SUMA - MMARG[INDICE])/(M - MMARG[INDICE]);
   END; (* LAMBDA *)
FUNCTION LAMBDA1(SUMA,M:REAL;MMARG:TMMARG):REAL;
   VAR
      SMAXMARG:REAL;
   BEGIN (* LAMBDA1 *)
      SMAXMARG:=MMARG[1] + MMARG[2];
      LAMBDA1:=(SUMA - SMAXMARG)/(2*M - SMAXMARG);
   END; (* LAMBDA1 *)
PROCEDURE ESCRMED3(LAMB:REAL;OPCION2:CHAR;BANDERA:BOOLEAN;CONT:INTEGER);
   VAR
      PREDICE:STRING[35];
      MI2,
      LPREDICE:INTEGER;
   BEGIN (* ESCRMED3 *)
      IF CONT=1
         THEN
            BEGIN (* THEN *)
               MI2:=MI-5;
               WRITELN(ARCHSAL,TIT3);
               WRITELN(ARCHSAL);
               WRITELN(ARCHSAL,BL:MI2,TO);
            END; (* THEN *)
      CASE OPCION2 OF
                  'A':PREDICE:=T1;
                  'B':PREDICE:=T2;
                  'C':PREDICE:=T3;
         END; (* CASE OPCION2 *)
      LPREDICE:=36-LENGTH(PREDICE);
      IF BANDERA=TRUE
         THEN
            WRITELN(ARCHSAL,BL:MI,PREDICE,BL:LPREDICE,LAMB:6:4)
         ELSE
            WRITELN(ARCHSAL,BL:MI,PREDICE,BL:LPREDICE,'INDEFINIDO');
   END; (* ESCRMED3 *)
BEGIN (* RPE *)
   CONT:=0;
   WRITELN('QUE INDICE DEBO CALCULAR ?');
   WRITELN;
   WRITELN(TO);
   WRITELN(BL:MI,'A. ',T1);
```

```
        WRITELN(BL:MI,'B. ',T2);
        WRITELN(BL:MI,'C. ',T3);
        REPEAT
           READLN(OPCION2);
           CASE OPCION2 OF
                        'A':BEGIN (* A *)
                              INDICE:=2;
                              MMARG[INDICE]:=MAXMARG(INDICE,NUMNIV,MARG);
                              IF MMARG[INDICE] <> M
                                 THEN
                                    BEGIN (* THEN *)
                                       BANDERA:=TRUE;
                                       MAXRENG(MRENG,NUMNIV);
                                       SUMA:=SUM(MRENG,MCOLUMN,NUMNIV,INDICE);
                                       LAMB:=LAMBDA(SUMA,M,INDICE,MMARG);
                                    END (* THEN *)
                                 ELSE
                                    BANDERA:=FALSE;
                            END; (* A *)
                        'B':BEGIN (* B *)
                              INDICE:=1;
                              MMARG[INDICE]:=MAXMARG(INDICE,NUMNIV,MARG);
                              IF MMARG[INDICE] <> M
                                 THEN
                                    BEGIN (* THEN *)
                                       BANDERA:=TRUE;
                                       MAXCOLUMN(MCOLUMN,NUMNIV);
                                       SUMA:=SUM(MRENG,MCOLUMN,NUMNIV,INDICE);
                                       LAMB:=LAMBDA(SUMA,M,INDICE,MMARG);
                                    END (* THEN *)
                                 ELSE
                                    BANDERA:=FALSE;
                            END; (* B *)
                        'C': BEGIN (* C *)
                               FOR J:=1 TO 2 DO
                                  BEGIN (* FOR J *)
                                     INDICE:=J;
                                     MMARG[INDICE]:=MAXMARG(INDICE,NUMNIV,MARG);
                                  END; (* FOR J *)
                               IF (MMARG[1] + MMARG[2]) <> (2*M)
                                  THEN
                                     BEGIN (* THEN *)
                                        BANDERA:=TRUE;
                                        INDICE:=1;
                                        MAXCOLUMN(MCOLUMN,NUMNIV);
                                        SUMA:=SUM(MRENG,MCOLUMN,NUMNIV,INDICE);
                                        INDICE:=2;
                                        MAXRENG(MRENG,NUMNIV);
                                        SUMA:=SUMA + SUM(MRENG,MCOLUMN,NUMNIV,INDICE);
                                        LAMB:=LAMBDA1(SUMA,M,MMARG);
```

```
                                    END (* THEN *)
                                ELSE
                                    BANDERA:=FALSE;
                            END; (* C *)
                    END; (* CASE OPCION2 *)
                CONT:=CONT + 1;
                ESCRMED3(LAMB,OPCION2,BANDERA,CONT);
                WRITELN('QUIRES QUE CALCULE OTRO INDICE DE RPE ? (SI=1/NO=0)');
                READLN(K);
                IF K=1
                   THEN
                      WRITELN('CUAL ? A,B o C')
            UNTIL K=0;
            WRITELN(ARCHSAL,BL:264);
        END; (* RPE *)
    PROCEDURE CONFIAB(MARG:TMARG;M:REAL;MENOR:INTEGER);
        CONST
            TIT4='MEDIDAS DE CONFIABILIDAD.';
            AT='LA KAPPA SIN PONDERAR.';
            BT='LA LAMBDA DE GOODMAN Y KRUSKAL.';
        TYPE
            TMEDID=ARRAY[1..2] OF REAL;
            TBAND=ARRAY[1..2] OF BOOLEAN;
        VAR
            MEDID:TMEDID;
            OPCION3:CHAR;
            FDIAG:REAL;
            BAND:TBAND;
        FUNCTION FRECDIAG(MENOR:INTEGER):REAL;
            VAR
                SUMFDIAG:REAL;
            BEGIN (* FRECDIAG *)
                CLOSE(ARCHENT);
                RESET(ARCHENT,'05:FREC.DATA.TEXT');
                READLN(ARCHENT);
                READLN(ARCHENT,FREC);
                SUMFDIAG:=FREC;
                FOR I:=1 TO MENOR DO
                    BEGIN (* FOR I *)
                        FOR J:=1 TO I DO
                            READ(ARCHENT,FREC);
                        READLN(ARCHENT,FREC);
                        SUMFDIAG:=SUMFDIAG + FREC;
                    END; (* FOR I *)
                FRECDIAG:=SUMFDIAG;
            END; (* FRECDIAG *)
        PROCEDURE KAP(VAR MEDID:TMEDID;VAR BAND:TBAND;MARG:TMARG;FDIAG,M:REAL;MENOR:INTEGER);
            VAR
                PMARG:REAL;
            FUNCTION PRODMARG(MENOR:INTEGER;MARG:TMARG):REAL;
```

```
        VAR
            SUMPMARG:REAL;
        BEGIN (* PRODMARG *)
            SUMPMARG:=0;
            FOR I:=1 TO MENOR + 1 DO
                SUMPMARG:=SUMPMARG + MARG[1,I]*MARG[2,I];
            PRODMARG:=SUMPMARG;
        END; (* PRODMARG *)
    FUNCTION KAPPA(PMARG,FDIAG,M:REAL):REAL;
        BEGIN (* KAPPA *)
            KAPPA:=(M*FDIAG - PMARG)/(M*M - PMARG);
        END; (* KAPPA *)
    BEGIN (* KAP *)
        PMARG:=PRODMARG(MENOR,MARG);
        IF PMARG <> (M*M)
            THEN
                BEGIN (* THEN *)
                    BAND[1]:=TRUE;
                    MEDID[1]:=KAPPA(PMARG,FDIAG,M);
                END (* THEN *)
            ELSE
                BAND[1]:=FALSE;
    END; (* KAP *)
PROCEDURE CLAMB(VAR MEDID:TMEDID;VAR BAND:TBAND;MARG:TMARG;FDIAG,M:REAL;MENOR:INTEGER);
    VAR
        MODA:REAL;
    FUNCTION MODAL(MENOR:INTEGER;MARG:TMARG):REAL;
        VAR
            MODTEMP:REAL;
        BEGIN (* MODAL *)
            MODTEMP:=MARG[1,1] + MARG[2,1];
            FOR I:=2 TO MENOR + 1 DO
                IF (MARG[1,I] + MARG[2,I]) > MODTEMP
                    THEN
                        MODTEMP:=MARG[1,I] + MARG[2,I];
            MODAL:=MODTEMP;
        END; (* MODAL *)
    FUNCTION CLAMBDA(MODA,FDIAG,M:REAL):REAL;
        BEGIN (* CLAMBDA *)
            CLAMBDA:=(2*FDIAG - MODA)/(2*M - MODA);
        END; (* CLAMBDA *)
    BEGIN (* CLAMB *)
        MODA:=MODAL(MENOR,MARG);
        IF MODA <> (2*M)
            THEN
                BEGIN (* THEN *)
                    BAND[2]:=TRUE;
                    MEDID[2]:=CLAMBDA(MODA,FDIAG,M);
                END (* THEN *)
```

```
            ELSE
               BAND[2]:=FALSE;
   END; (* CLAMB *)
PROCEDURE ESCRMED4(MEDID:TMEDID;BAND:TBAND;OPCION3:CHAR);
   TYPE
      TNOMED=ARRAY[1..2] OF STRING[31];
   VAR
      NOMED:TNOMED;
      LNOMED,
      TOPE1,
      TOPE2:INTEGER;
   BEGIN (* ESCRMED4 *)
      NOMED[1]:=AT;
      NOMED[2]:=BT;
      CASE OPCION3 OF
                  'A':BEGIN (* A *)
                        TOPE1:=1;
                        TOPE2:=1;
                      END; (* A *)
                  'B':BEGIN (* B *)
                        TOPE1:=2;
                        TOPE2:=2;
                      END; (* B *)
                  'C':BEGIN (* C *)
                        TOPE1:=1;
                        TOPE2:=2;
                      END; (* C *)
          END; (* CASE OPCION3 *)
      FOR I:=TOPE1 TO TOPE2 DO
          BEGIN (* FOR I *)
            LNOMED:=32-LENGTH(NOMED[I]);
            IF BAND[I]=TRUE
               THEN
                  WRITELN(ARCHSAL,BL:MI,NOMED[I],BL:LNOMED,MEDID[I]:6:4)
               ELSE
                  WRITELN(ARCHSAL,BL:MI,NOMED[I],BL:LNOMED,'INDEFINIDO');
          END; (* FOR I *)
      WRITELN(ARCHSAL,BL:264);
   END; (* ESCRMED4 *)
BEGIN (* CONFIAB *)
   WRITELN('QUE INDICE(S) CALCULO ?');
   WRITELN(BL:MI,'A. ',AT);
   WRITELN(BL:MI,'B. ',BT);
   WRITELN(BL:MI,'C. AMBOS.');
   READLN(OPCION3);
   FDIAG:=FRECDIAG(MENOR);
   CASE OPCION3 OF
               'A':KAP(MEDID,BAND,MARG,FDIAG,M,MENOR);
               'B':CLAMB(MEDID,BAND,MARG,FDIAG,M,MENOR);
```

```
                        'C':BEGIN (* C *)
                                KAP(MEDID,BAND,MARG,FDIAG,M,MENOR);
                                CLAMB(MEDID,BAND,MARG,FDIAG,M,MENOR);
                            END; (* C *)
                END; (* CASE OPCION3 *)
            WRITELN(ARCHSAL,TIT4);
            WRITELN(ARCHSAL);
            ESCRMED4(MEDID,BAND,OPCION3);
            WRITELN(ARCHSAL,BL:264);
        END; (* CONFIAB *)
    PROCEDURE INDICES(NUMNIV:TNUMNIV);
        CONST
            T='MEDIDAS BASADAS EN ';
            G1='LA ESTADISTICA JI-CUADRADA PARA PROBAR INDEPENDENCIA.';
            G2='LA LOGICA DE REDUCCION PROPORCIONAL EN EL ERROR.';
            G3='MEDIDAS DE CONFIABILIDAD (TABLAS CUADRADAS).';
        VAR
            OPCION:CHAR;
        BEGIN (* INDICES *)
            WRITE('ESTE PROGRAMA CALCULA ALGUNOS INDICES PERTENECIENTES ');
            WRITELN('A LOS SIGUIENTES GRUPOS:');
            WRITELN;
            WRITELN('A. ',T,G1);
            WRITELN('B. ',T,G2);
            WRITELN('C. ',G3);
            WRITELN;
            REPEAT
                WRITELN('A QUE GRUPO PERTENECE LA MEDIDA QUE DEBO CALCULAR ?');
                READLN(OPCION);
                CASE OPCION OF
                        'A':JICUAD(NUMNIV,MARG,M,MENOR);
                        'B':RPE(NUMNIV,M,MARG);
                        'C':BEGIN (* C *)
                                IF NUMNIV[0]=NUMNIV[1]
                                    THEN
                                        CONFIAB(MARG,M,MENOR)
                                    ELSE
                                        WRITELN('LA TABLA NO ES CUADRADA.');
                            END; (* C *)
                END; (* CASE OPCION *)
            WRITELN('QUIERES QUE CALCULE OTRO INDICE ');
            WRITELN(BL:5,'A. BASADO EN ',G1);
            WRITELN(BL:5,'B. BASADO EN ',G2);
            WRITELN(BL:5,'C. DE CONFIABILIDAD ?');
            WRITELN('SI=1 ; NO=0');
            READLN(K)
```

```
        UNTIL K=0;
    END; (* INDICES *)
BEGIN (* PROGRAMA PRINCIPAL *)
    RESET(ARCHENT,'#5:FREC.DATA.TEXT');
    REWRITE(ARCHSAL,'#5:RESULT2.TEXT');
    WRITELN(BL:120);
    WRITELN(BL:14,'MEDIDAS DE ASOCIACION NOMINALES.');
    WRITELN(BL:25,'PARTE 2');
    WRITELN;
    LEE(M,NUMNIV);
    MAXMIN(NUMNIV,MENOR,MAYOR);
    MARGIN(NUMNIV,MARG,MENOR,MAYOR);
    INDICES(NUMNIV);
    CLOSE(ARCHENT);
    CLOSE(ARCHSAL,LOCK);
    WRITELN('FIN DE LA PARTE 2.');
END.  (* PROGRAMA PRINCIPAL *)
```