COEFICIENTES DE ASOCIACIÓN
(Guía de clase)

FICHA Nº 13

A. Cuando el interés teórico se dirige a conocer si un fenómeno cualquiera está definido simultáneamente por distintas variables, la estadística pone a nuestra disposición desde principios del siglo XX distintas técnicas para establecer si existe o no asociación o correlación entre las variables respeto a un universo de análisis.

✎ Originalmente, los términos asociación y correlación hacían referencia a dos ramas contrapuestas de la estadística: la paramétrica y la no paramétrica. Esto lleva a que los coeficientes sean clasificados también según estas dos ramas:

i) Coeficiente de asociación es el nombre que reciben las medidas de relación para las variables no paramétricas (nominales y ordinales). Una excepción a esta "reserva de nombres" la representa el libro de Siegel & Castellan (2003)

ii) Coeficientes de correlación se reservan para las variables métricas.

✎ Según Cortés & Rubalcaba (1987), la asociación entre dos variables puede definirse de dos formas.

i) La más simple y frecuente forma de examinar la asociación es por contraposición a la *independencia estadística.* Para esto se desarrollan coeficientes que tratan de resumir la distribución conjunta a través de magnitudes rápidamente interpretables.

- Es de observarse en este caso que la hipótesis nula se restringe a sostener que existe independencia entre las variables, aunque la hipótesis sustantiva sea muy sofisticada.

ii) La segunda forma de hablar de asociación es verificando si existe o no *articulación entre las proposiciones deducidas desde el marco teórico (incluidos los antecedentes) y las distribuciones observadas empíricamente.*

✎ En el análisis de asociación o correlación entre dos variables es necesario distinguir 4 aspectos distintos que pueden ser objetivo de una hipótesis.

i) La **existencia** (si/no) de una asociación simplemente verificada a través del rechazo de la hipótesis nula que afirma independencia estadística.

ii) La **magnitud** de la asociación existente entre dos variables con base en una

escala fácilmente interpretable (muy fuerte/ fuerte/moderada/ débil/ despreciable) y que permita comparar distintas asociaciones estimadas en distintos cruces. Esta noción de magnitud de asociación supone que antes se ha descartado la hipótesis nula de la independencia estadística.

iii) El **sentido** de la asociación entre las variables distribuidas conjuntamente. Por ejemplo, afirmando que cuando una se incrementa otra disminuye. Es claro que esto es posible con variables cuya escala de medición es al menos ordinal.

iv) La **forma** de la relación entre las variables examinadas. Si se imagina una representación gráfica, se puede esperar que dos variables puedan tener una relación lineal, curvilineal, exponencial, logística, etc. Este es un punto importante porque algunos coeficientes están diseñados para capturar únicamente relaciones lineales.

Objetivo del análisis	La medida de asociación indicará:	Otras propiedades de la medida:	Tipo de variables
Existencia	si existe o no asociación		nominales, ordinales, intervales, de razón
Magnitud	un valor que permita fácilmente ver qué tan fuerte es la relación	Podría ser de utilidad que hubiera un único valor mínimo y un valor máximo (por, ejemplo 0 y 1)	nominales, ordinales, intervales de razón.
Sentido	si la relación es directa o inversa	el coeficiente tendrá "signo" (+) ó (-)	ordinales, intervales, de razón
Forma	Al menos debería poder captar formas no-lineales de asociación	Lo más útil aquí será graficar la relación en ejes cartesianos	intervales, de razón

v) Lógicamente, el esquema expone que las medidas que permiten analizar la forma de la relación, incorporan todas las propiedades de las medidas que permiten analizar el sentido, la magnitud y la existencia. Lo mismo para cada una de las anteriores medidas.

vi) Por una razón de "economía de esfuerzos", si una hipótesis sólo refiere a la existencia de una asociación, será condición suficiente contar con una medida que indique si existe o no asociación.

vii) Resulta también lógico que la posiblidad de realizar alguno de los cuatro objetivos del análisis de asociación está restringida por el tipo de variables con que se cuenta. La división más clara está entre las variables nominales y ordinales a las cuales sólo se pueden aplicar medidas de existencia y magnitud y las restantes.

viii) Nuevamente son una excepción las variables dicotómicas, las cuales pueden ser

objeto de análisis de sentido de asociación.

ix) Desde el punto de vista estadístico, se suele sostener que los coeficientes más apropiados por versátiles y elocuentes, tienen las siguientes propiedades:

- No están afectados por el N total de la distribución. Es decir están normalizados.
- El valor "cero" indica la inexistencia de una relación o más estrictamente, independencia estadística.
- Existe un valor máximo teóricamente establecido, el mismo para cualquier distribución, que indica una relación de asociación o correlación perfecta.
- En el caso de que el objetivo sea un análisis del sentido, también existirá un valor teórico mínimo menor que cero que indicará asociación perfecta pero en sentido inverso.
- Los estadísticos de asociación que cumplen con estas propiedades varían entre 0 y 1; o entre -1, 0 y 1.

✎ En el análisis de asociación o correlación se suceden cuatro etapas distintas igualmente importantes: la etapa lógica de formulación de la proposición empírica; la etapa de cálculo de la medida de asociación; la etapa inferencial y finalmente, la etapa de interpretación de los resultados.

i) La proposición empírica ha de ser expresada en forma tal que se pueda esclarecer directamente cuál de los cuatro objetivos anteriores (existencia, magnitud, sentido y forma) es el objetivo de la hipótesis. En virtud de esto se plantea la hipótesis nula en todos los casos estará enunciando la independencia estadística.

ii) la etapa de **cálculo** es específica para cada coeficiente de asociación que se trabaje y puede ser más o menos directo.

iii) la etapa **inferencial** tiene su orginen en los siguientes reconocimientos:

- estar trabajando con una muestra de la población
- establecer un nivel de error para la aceptación o rechazo de la hipótesis nula
- contrastar el resultado obtenido empíricamente con el resultado que se debiera observar en una situación en la cual no existe asociación.

iv) la etapa de **interpretación** consiste en la contrastación de los resultados obtenidos con base en la distribucion conjunta observada con la distribución esperada según la o las hipótesis de partida.

B. Conviene explicitar una primera noción de **independencia estadística** dada su importancia para definir la asociación en el caso de un cuadro y derivar de ella un primer coeficiente que permite establecer si existe o no asociación.

✎ Es importante volver a recordar que la hipótesis nula que se somete en todos los análisis de asociación basados en el enfoque de independencia, es la independencia estadística entre las variables.

i) Esta hipótesis debe ser establecida en la etapa lógica del análisis, donde se contrapone la hipótesis nula a la(s) proposiciones empíricas.

ii) Subyace a los cálculos que se realicen de las medidas de asociación seleccionadas para trabajar sobre el cuadro.

iii) Finalmente se contrasta en la etapa inferencial a través de pruebas estadísticas específicas que permiten decidir si lo observado se aleja en una magnitud suficientemente razonable de lo esperado como para descartar que las diferencias se deben al azar.

✎ Se afirmará que dos variables son estadísticamente independientes cuando en **cada celda** de un cuadro bivariado (j,m) se han observado la misma cantidad de frecuencias absolutas ($N_{j,m}$) que las que se esperaban ($E_{j,m}$) en una situación de distribución al azar.

i) En un cuadro bivariado construido según la convención, con la variable *independiente* en las columnas y la variable *dependiente* en las filas, se adopta la siguiente nomenclatura:

	Categoría 1	Categoría 2	Categoría ...	Categoría J	Total filas
Categoría 1	$N_{1,1}$	$N_{2,1}$	$N_{...,1}$	$N_{j,1}$	R_i
Categoría 2	$N_{1,2}$	$N_{2,2}$	$N_{...,2}$	$N_{j,2}$	R_2
Categoría ...	$N_{1,..}$	$N_{2,...}$	$N_{....,..}$	$N_{j,..}$	$R_{..}$
Categoría m	$N_{j,m}$	$N_{2,m}$	$N_{...,m}$	$N_{j,m}$	R_m
Total columnas	C_1	C_2	$C_{...}$	C_j	N

ii) Las frecuencias absolutas se han representado con la letra N, los totales de casos

para cada columna con la letra C, los totales de casos para las filas con la letra R. Los subíndice primero indican la categoría respectiva de la variable independiente (columna en este caso *convencional*) y luego la categoría de la variable dependiente.

iii) El cálculo de las frecuencias esperadas para la celda $N_{1,1}$ se realiza de la siguiente forma:

[I.1] $E_{1,1} = (C_1 * R_1) / N$

En términos verbales, las frecuencias **esperadas** (E)para cualquier celda (j,m) es igual al producto del marginal de columna por el marginal de la fila dividido entre el número total de casos de la tabla.

iv) Debe observarse que:

- por lo general, el número de frecuencias **esperadas** (E) en una celda será una fracción y no un número natural.
- En ningún caso es igual a 0
- La suma de todas las frecuencias esperadas (E) s es igual a N

✎ Es posible que en la generalidad de la situaciones, la mera comparación entre las frecuencias esperadas y las observadas de cada celda no permita establecer directamente si existe asociación o no entre dos variables. En consecuencia, será necesario contar con una medida que informe de este aspecto.

✎ El **coeficiente de Chi (o ji) cuadrada, simbólicamente** χ^2 es una medida o coeficiente que permite contrastar la hipótesis de que dos variables distribuidas conjuntamente en un cuadro son estadísticamente independientes[1]. Su procedimiento de cálculo es el siguiente:

i) Para cada celda hay que calcular las "frecuencias esperadas". Supongamos que lo

[1] Es de notarse estrictamente no son sinóminos los términos Ji cuadrada y χ^2: el primero es una función de los datos (estadístico); la segunda es una distribución teórica en el muestreo (tema de la estadística inferencial). Siegel (2003) plantea la distinción de que el estadístico calculado en una distribución conjunta cualquiera sigue una distribución en el muestreo que se aproxima a una distribución de χ^2 conforme se incrementa el número de casos hasta el infinito ("asintóticamente").

hacemos para la primera celda del cuadro (columna 1, fila 1).

ii) Una vez que tenemos todas frecuencias esperadas ($E_{j,m}$) para todas las celdas del cuadro, conviene realizar una prueba para comprobar de que no hubo errores de cálculo. Esto se hace simplemente, sumando todas las "frecuencias esperadas", las cuales deben ser igual al total de casos del cuadro es decir, el "N". Dado que hemos trabajado con decimales y con redondeos, es probable que los números no sean exactamente iguales. Si el resultado no es aproximadamente igual, sería conveniente revisar.

iii) El siguiente paso consiste en calcular para cada celda del cuadro la discrepancia entre lo esperado y lo observado. Esto se hace simplemente restando ambos números. Pero aquí es necesario hacer dos correcciones.

- La primera es elevar al cuadrado las diferencias calculadas en cada celda. Esto se hace para eliminar los signos; si no lo hicieramos, las diferencias terminarían por anularse.

- La segunda corrección es dividir el cuadrado calculado en cada celda entre las "frecuencias esperadas" en esa celda. Esto se llama "normalización" y el objetivo es controlar el hecho de que las celdas tienen diferentes cantidades de casos.

iv) Una vez que tenemos estos cuadrados, estamos en condiciones de sumar todos los valores. El resultado va a ser el valor de la Chi cuadrada para nuestro cuadro.

C. En términos más generales, un **coeficiente** es un modelo matemático construido para resumir las propiedades de una relación sea con respecto a la independencia estadística, sea respecto a una proposición pre-establecida.

✎ Todos los coeficientes tienen supuestos más o menos "sofisticados" que requieren ser examinados para su aplicación , tanto desde el un punto de vista teórico de su adecuación a la distribución esperada, como desde un punto de vista estadístico. Esquemáticamente:

i) El supuesto teórico de un coeficiente relaciona las operaciones de cálculo que se realizan sobre la matriz de datos con la proposición empírica que se ha planteado.

- ¿Cómo se espera que se distribuyan los datos entre las celdas del cuadro de acuerdo a lo establecido en la hipótesis?

- En la fórmula del coeficiente: ¿a qué celdas se le está dando más peso?

ii) El supuesto estadístico tiene que ver con el nivel de medición de las variables distribuidas conjuntamente y con el número de categorías de cada variable.

- La primera restricción en el conjunto de coeficientes elegibles la impone la variable de menor nivel de medición. Habrán coeficientes para variables nominales, ordinales.

- Una segunda restricción la impone el más alto número de categorías de las variables distribuidas: habrán coeficientes para variables dicotómicas, tricotómicas, etc.

iii) El supuesto lógico tiene refiere a si la proposición (teórica o empírica) establece alguna dirección causal en el análisis.

- Hay coeficientes que se denominan simétricos porque no suponen dirección causal.

- Los coeficientes que suponen causalidad se llaman asimétricos.

✎ En esta ocasión nos interesarán únicamente los coeficientes de asociación que conforman la estadística paramétrica y examinaremos el caso en que la asociación es definida como oposición a la independencia.

D. Para el análisis de una distribución de dos dicotomías, los coeficientes más utilizados son la ϕ y la Q.

✎ Estos coeficientes tienen algunas propiedades comunes de interés:

i) Están normalizados: sus magnitudes no dependen del tamaño de la tabla.

ii) Son altamente sensibles a la distribución empíricamente observada, traduciendo concentraciones de los casos en algunas celdas en magnitudes.

iii) Tienen un recorrido teórico cerrado entre -1 y 1, indicando situaciones de asociación perfecta y de independencia estadística.

✎ Estos coeficientes se diferencian en la "sensibilidad rinconal":

i) El coeficiente Q es altamente sensible a la existencia de una celda que en términos relativos se "está vaciando". Su valor máximo se alcanza cuando en una celda no hay ningún caso: esto es lo que se conoce como "sensibilidad rinconal".

ii) El coeficiente ϕ alcanza su valor máximo sólo cuando una de las dos diagonales se ha vaciado.

COEFICIENTES DE ASOCIACIÓN
(Guía de clase)

✎ Las fórmulas de cálculo son las siguientes:

Phi:

$$\mathbf{Phi} = \frac{(a*d)-(b*c)}{\sqrt{(a+b)*(a+c)*(b+d)*(c+d)}}$$

Q de Yule:

$$Q = \frac{(a*d)-(b*c)}{(a*d)+(b*c)}$$

E. En la siguiente tabla se pueden observar un listado de algunos coeficientes.

Niveles de medición de las dos variables		**Posibilidades**	**Coeficientes de Asociación**
Ambas dicotómicas		Existencia	Ji cuadrado (χ^2)
		Existencia y magnitud	V de Cramer
		Existencia, magnitud y sentido	Phi (ϕ) Q de Yule
Ambas nominales	pluri-cotómicas "(r * c)"	Existencia	Ji cuadrado (χ^2) Cuadrático cuadrado medio (ϕ^2)
		Existencia y magnitud	Contingencia de Tschuprow (T^2) V de Cramer y (C^2) de Cramer Contingencia de Pearsons (P^2)
Ambas ordinales	varias categorías	Existencia, magnitud y dirección	Rho de Spearman, Tau-C (τ-c) Gamma (γ)
Una nominal - otra interval	Existencia y magnitud		Coeficiente Etha
una dicotómica - otra interval	Existencia, magnitud, dirección y forma		Coeficiente r de Pearson

COEFICIENTES DE ASOCIACIÓN
(Guía de clase)

F. La situación más simple de ejemplificar es el análisis de relaciones entre variables dicotómicas representadas en cuadros de 2 x 2 donde los coeficientes que más frecuentemente se utilizan son ϕ y Q.

✎ Primero, se examina la proposición empírica para indagar lógicamente cuál es la distribución esperada de los datos.

i) *Una discusión importante dentro de las políticas educativas tiene que ver con el efectivo cumplimento de la escolarización universal en los niveles que la sociedad ha definido como obligatoria. Por lo general, esto se informa a través de la tasa escolaridad neta para la cohorte etaria comprendida dentro de las normas constitucionales y legales. Sin embargo, la investigación ha mostrado reiteradamente que existen fenómenos de exclusión educativa relacionados con el hecho de que un hogar sea pobre. Entre éstos, los niños son enviados más tardíamente a escuela y son retirados más tempranamente una vez que los padres consideran que se han cumplido las metas mínimas del aprendizaje de la lectura y la escritura. Para el caso de Chiapas en los cuatro municipios seleccionados, se conoce que para la cohorte de niños entre 8 y 12 años censados en el año 2000, la tasa de escolarización era del 93,1%. Se supone que esta tasa desciende en forma significativa entre los niños que viven en hogares pobres y sube tambiénen forma importante para los que no viven en hogares pobres.*

ii) Según lo enunciado, se propone la siguiente distribución esperada, utilizando como criterio de pobreza la Línea de Pobreza 1 del Comité Técnico. Las celdas rayadas representan la mayor concentración de casos.

	No- Pobres	Pobres
No asistentes		
Asistentes		

iv) Se espera que la presencia de un atributo implique la ausencia del otro. Esto significa que los casos tenderán a concentrase en la diagonal secundaria o menor del cuadro.

COEFICIENTES DE ASOCIACIÓN
(Guía de clase)

✎ Se delimitan los coeficientes de asociación a utilizarse de acuerdo al menor nivel de medición de las variables involucradas y al mayor número de categorías de las variables.

i) Las dos variables en juego son dicotómicas: indican la presencia o ausencia de atributos (pobreza, asistencia).

ii) Conforman una tabla cuadrada de 2 x 2.

ii) Los coeficientes que se pueden usar son todos aquellos que permitan realizar un análisis para tablas de 2 x 2. De preferencia se opta por ϕ y Q dada la relativa facilidad de los cálculos involucrados.

iii) Según el principio de isomorfía, se espera que los casos se concentren en la diagonal secundaria o menor, por lo que se deberá utilizar ϕ dadas sus propiedades estadístico.

✎ Se realizan los cálculos para el cuadro donde se ha observado la distribución empírica de las variables.

abla de contingencia EL NIÑO (8 A 12 AÑOS) ASISTE A LA ESCUEI NIVEL DE POBREZA 1

			NIVEL DE POBREZA 1		
			no pobres	pobres	Total
EL NIÑO (8 A 12 AÑOS) ASISTE A LA ESCUEL.1	no	Recuento	72069	68508	140577
		% de NIVEL DE POBREZA 1	39,7%	45,9%	42,5%
	sí	Recuento	109274	80885	190159
		% de NIVEL DE POBREZA 1	60,3%	54,1%	57,5%
Total		Recuento	181343	149393	330736
		% de NIVEL DE POBREZA 1	100,0%	100,0%	100,0%

[I.2]

$$\mathbf{F} = \frac{(a*d)-(b*c)}{\sqrt{(a+b)*(a+c)*(b+d)*(c+d)}}$$

[I.3] $\phi =$ *[(72069 *80885) - (68508 * 109274)]*
SQR (140577 190159* 181343 * 149393)*

[I.4] ϕ = *-0,062*

G. Las **fórmulas** de cálculo para otros coeficientes son las siguientes:

Coeficiente de asociación C^2 de Cramer:

$$C^2 = \frac{\chi^2}{N(\min.r,c)}$$

Coeficiente de asociación de V de Cramer en Siegel (2003) y en el SPSS:

$$V = \sqrt{C^2}$$

Coeficientes de Tau:

$$\tau_a = \frac{fc - fi}{N(N-1)/2}$$

$$\tau_b = \frac{f_c - f_i}{\sqrt{(fc + fc + E_x)(fc + fi + E_y)}}$$

Gamma:

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

donde:

$$fc = A(E + F + H + I) + B(F + I) + D(H + I) + E(I)$$

$$fi = C(E + H + D + G) + B(D + G) + F(H + G) + E(G)$$

Para el cálculo de este coeficiente se adopta el criterio de nombrar las celdas de un cuadro de la siguiente forma:

A	B	C
D	E	F
G	H	I

Coeficiente Etha:

$$Etha = \frac{\sum_{j=1} nj\left(\overline{Y_j} - \overline{Y}\right)^2}{\sum \left(Y_i - \overline{Y}\right)^2}$$

donde:

- n_j es el número de observaciones del sub-grupo j yx
- Y_j media del grupo j.
- Y^j media aritmética total.