UNIÓN
REVISTA IBEROAMERICANA DE EDUCACIÓN MATEMÁTICA

www.fisem.org/web/union

ISSN: 1815-0640
Número 34. Junio de 2013
páginas 85-96

Problemática y recursos en la interpretación de las tablas de contingencia

Gustavo R. Cañadas, José M. Contreras, Pedro Arteaga, María M. Gea

Fecha de recepción: 13/12/2011
Fecha de aceptación: 20/12/2012

Resumen	Las tablas de contingencia aparecen diariamente en la prensa y el trabajo profesional, pero se les presta poca atención en la enseñanza. En este trabajo resumimos las investigaciones sobre errores en su lectura y describimos algunas medidas de asociación que pueden ayudar a interpretar correctamente la asociación de las variables en dicha tabla y podrían ser incluidos al final de la Educación Secundaria Obligatoria o del Bachillerato. **Palabras clave**: Tablas de contingencia, asociación y dificultades de interpretación.
Abstract	Contingency tables are frequent in daily press and professional work, but receive little attention in education. In this paper we summarize research on reading errors in these tables and describe some association coefficients that can help interpret correctly the association between variables represented in a contingency table. These tolos may be included at the end of Secondary Education or in High School levels. **Keywords**: Contingency table, association and interpretation difficulties.
Resumo	Las Tabelas de contingência aparecem diariamente nos jornais e trabalho profissional, mas recebem pouca atenção na educação. Este artigo resume a pesquisa sobre erros de leitura e descrevem algumas medidas de associação que podem ajudar a interpretar corretamente a associação das variáveis na tabela e poderian ser incluídas no final do ensino secundário obrigatório ou ensino médio. **Palavras-chave:** Tabelas de contingência, associação e dificuldades de interpretação.

1. Introducción

En la actualidad, cada vez es más habitual valorar la importancia de analizar críticamente la información estadística, como herramienta valiosa para conocer y analizar mejor la realidad. Una de las formas más comunes de presentación de la información en los medios de comunicación ó incluso Internet es mediante el formato de tabla de doble entrada o tabla de contingencia, tema al que se presta poca atención en la enseñanza, suponiendo que su lectura e interpretación es sencilla.

Las tablas se pueden utilizar para comunicar información y como instrumento de análisis de datos, así como para retener en la memoria una gran cantidad de información en forma eficiente (Cazorla, 2002). Tienen un papel esencial en la organización, descripción y análisis de datos, al ser un instrumento de

transnumeración, forma básica de razonamiento estadístico que proporciona nueva información, al cambiar de un sistema de representación a otro (Wild y Pfannkuch, 1999). En la enseñanza, este formato de presentación de la información ayuda a visualizar conceptos y relaciones abstractas difíciles de comprender (Postigo y Pozo, 2000). Sin embargo, a pesar de su supuesta simplicidad no es un tema tan sencillo como aparenta, incluso en el caso elemental en que la tabla sólo tiene dos filas y dos columnas (tabla 2x2).

En este trabajo se hace un breve resumen de los resultados de la investigación en Psicología y Didáctica de la Matemática sobre el tema. Seguidamente se describen algunos procedimientos sencillos que ayudan a identificar el tipo de asociación en las tablas 2x2. La finalidad es informar al profesor de Educación Secundaria y Bachillerato, quien podría incluir el estudio de estos procedimientos en la enseñanza de la estadística.

2. Dificultades de interpretación en las tablas de contingencia

Estas investigaciones las inician Inhelder y Piaget (1955), quienes pensaban que la comprensión de la asociación sería el último paso en el desarrollo del razonamiento sobre probabilidad. Los autores describen las estrategias en los juicios de asociación siguiendo el esquema de la Tabla 1. Cuando se pide estudiar la posible asociación entre las variables *A* y *B* a partir de los datos de la tabla, los sujetos al alcanzar la adolescencia comienzan usando sólo la celda *a* para juzgar la asociación; es decir consideran que hay asociación sólo si el número de casos en que se presentan a la vez *A* y *B* es suficientemente elevado.

	B	**No B**	**Total**
A	**A**	**B**	**a + b**
No A	**C**	**D**	**c + d**
Total	**a + c**	**b + d**	**a + b + c + d**

Tabla 1. Esquema de una tabla de contingencia 2x2

Entre los 12-15 años, los alumnos solamente comparan celdas dos a dos (por ejemplo comparan a con *b*), sin entender que *a* y *d* tienen el mismo peso en relación a la asociación. Otro nivel posterior sería comprender cuales son los casos favorables (*a* y *d*) y desfavorables (*b* y *c*) de la asociación, sin compararlos. Finalmente, se establecen las relaciones diagonales (*a* y *d* serían favorables a la asociación, mientras que *b* y *c* serían contrarias), comparándolas entre sí o con el total (*a*+*b*+*c*+*d*).

Jenkins y Ward (1965) indican que la estrategia de comparar las diagonales sólo se puede usar con frecuencias marginales iguales para la variable independiente y proponen para casos generales como estrategia correcta comparar la diferencia entre las probabilidades *P(B/A)* y *P(B/noA)*. Otros autores han estudiado la influencia de las teorías previas en el contexto del problema en los juicios de asociación (Jennings, Amabile y Ross, 1982; Wright y Murphy, 1984; Alloy y Tabachnik, 1984). En términos generales se puede decir, que cuando los datos no reflejan los resultados esperados por estas teorías, aparece en los sujetos un conflicto cognitivo. Más recientemente Estepa (1993) estudia las concepciones que muestran los sujetos respecto a la asociación, describiendo las siguientes:

1. *Concepción causal:* Este término se emplea cuando el sujeto sólo considera la asociación entre variables si puede adjudicarse a la presencia de una relación causal entre las mismas.
2. *Concepción determinista:* Este término describe el caso en que los sujetos no admiten el caso de excepciones, implicando esto que a cada valor de la variable independiente le corresponde un solo valor de la variable dependiente, es decir, esperan una dependencia funcional de tipo determinista. En el caso de la tabla 2x2, este caso se presenta cuando los sujetos afirman que no hay correspondencia por el hecho de que en la tabla hay casos en las celdas *b* o/y *c*. Otro ejemplo sería el caso en que el sujeto exige la existencia de una expresión algebraica que relacione las variables.
3. *Concepción unidireccional:* En este caso el estudiante no admite la asociación inversa, considerándose la intensidad de la asociación, pero no su signo. Presentándose casos en los que se considera la asociación inversa como independencia.
4. *Concepción local:* Esta concepción se presenta cuando los sujetos, dan su solución mirando únicamente algunos casos aislados. Por ejemplo, cuando sólo se tienen en cuenta los casos que confirman la asociación, observando una sola distribución condicional o fijándose en la celda de máxima frecuencia.

En lo que sigue utilizamos un ejemplo sencillo, para analizar posibles estrategias correctas para evaluar la asociación en una tabla 2x2 e introducimos algunas medidas de asociación, sencillas, que se podrían estudiar al final de la educación secundaria o Bachillerato para ayudar a los alumnos a establecer un juicio de asociación correcto en una tabla 2x2, superando los sesgos descritos.

3. Frecuencias en una tabla 2x2

La tabla de contingencia nos proporciona una forma resumida de representar datos de dos variables que se quieren estudiar, como vemos en el siguiente ejemplo.

Ejemplo 1. Problemas de una enfermedad. Supongamos que en un hospital se comparan dos fármacos, uno nuevo y uno antiguo, en 300 pacientes clasificados en dos tipos de niveles de una cierta enfermedad (altos y bajos). Tras la recogida de datos se obtienen los que se incluyen en la Tabla 2. ¿Cómo podemos ver si el tratamiento nuevo es preferible al anterior?

Tipo de tratamiento(X)	Problemas neuronales (Y)	
	Altos(y1)	Bajos(y2)
Antiguo(x1)	40	60
Nuevo(x2)	70	130

Tabla 2. Resultados de un estudio clínico

Para analizar esta situación se consideran dos variables: la variable *X* que representa el tratamiento, con dos valores (x_1, x_2), y la variable *Y*, que se refiere a los problemas (y_1, y_2). Usualmente f_{ij} indica la frecuencia absoluta de cada celda (con que aparece el par (x_i, y_j)) y h_{ij} la frecuencia relativa del par de valores (x_i, y_j), verificándose la relación siguiente.

$$h_{ij}=f_{ij}/n$$

Una posible representación gráfica de esta tabla sería el diagrama de barras apilado (Figura 1), que también podría representarse en porcentajes, bien absolutos (respecto al total de la tabla), bien relativos (respecto a cada valor de la variable *X*).

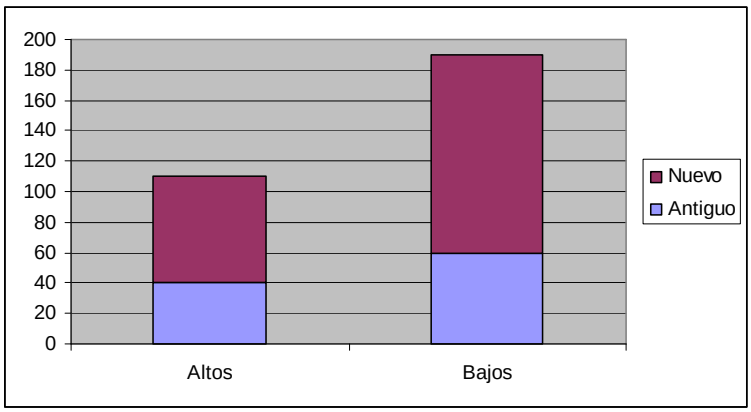

Figura 1. Datos del experimento

A partir de la tabla de contingencia (bidimensional) se pueden obtener diferentes distribuciones de una variable (unidimensional). Si en la tabla de frecuencias se suman las frecuencias por columnas, se obtiene en cada columna j, el número de individuos $f_{.j}$ con un valor de la variable $Y=y_j$, independientemente del valor *X*. A la distribución así obtenida se le conoce como distribución marginal de la variable *Y*. De forma análoga podemos definir la distribución marginal de la variable *X*. En el ejemplo hay 110 pacientes con problemas con un nivel alto y 190 pacientes con problemas con un nivel bajo, 100 pacientes con el tratamiento antiguo y 200 pacientes con el tratamiento nuevo, y 300 pacientes en total.

Es posible que el interés sea focalizar solamente en una parte de los pacientes, por ejemplo en los pacientes con el tratamiento nuevo. Para ello se calculan distribuciones condicionales. Si representamos por $h(x_i|y_j)$ la frecuencia relativa condicional del valor x_i entre los individuos que presentan el carácter y_j, obtenemos la tabla 3, donde observamos que el porcentaje de enfermos con problemas altos y bajos es aproximadamente el mismo (pero no exactamente) en los dos tratamientos y que las frecuencias condicionales por filas suman 1. Igualmente se calcularían las distribuciones condicionales por columna.

$$h(x_i|y_j) = \frac{f_{i,j}}{f_{.j}} = \frac{h_{i,j}}{h_{.j}}$$

Tipo de tratamiento(X)	Problemas neuronales (Y)		
	Altos(y1)	Bajos(y2)	Total
Antiguo(x1)	0,4	0,60	1
Nuevo(x2)	0,35	0,65	1

Tabla 3. Distribuciones condicionales por fila

4. El concepto de asociación

En general, cuanto tenemos una tabla de contingencia estamos interesados en ver si las variables en filas y columnas están relacionadas entre sí. Hablaremos en este caso de asociación, para diferenciar del concepto de correlación, que se refiere a variables numéricas (mientras que, generalmente las variables en una tabla de

contingencia son cualitativas). En lo que sigue tratamos de proporcionar instrumentos para contestar a las preguntas siguientes (Batanero y Díaz, 2008):

1. ¿Cómo podemos determinar si hay o no asociación entre las variables en una tabla de contingencia?
2. ¿Podría medir la intensidad de esta relación mediante un coeficiente (coeficiente de asociación)? ¿Se puede medir el signo en algunos casos?
3. ¿Cómo interpretar estos coeficientes?

a) Independencia total					b) Asociación parcial					c) Asociación perfecta				
	Y_1	Y_2	Y_3	T		Y_1	Y_2	Y_3	T		Y_1	Y_2	Y_3	T
X_1	10	20	70	100	X_1	10	80	10	100	X_1	100			100
X_2	20	40	140	200	X_2	80	20	100	200	X_3			200	200
T	30	60	210	300	T	90	100	110	300	X_3		100		100

Figura 2. Diferentes tipos de asociación (T=Total)

Para deducir si las dos variables que forman una tabla de contingencia están o no asociadas se requiere un proceso de cálculo a partir de las frecuencias de la tabla. Podemos encontrar una variedad de situaciones (ver Figura 2)

- En el caso c) observamos que a cada valor de la variable *X* corresponde un solo valor de la variable *Y* y viceversa. Es claro que en este caso las variables están asociadas, y más aún se trataría de una asociación perfecta pues con toda seguridad podemos predecir el valor que tomara *Y* sabiendo el valor de *X* y viceversa.
- En el caso a) independencia, observamos que las frecuencias absolutas dobles de los valores de Y_1, Y_2, Y_3 son proporcionales en X_1 y X_2. Es decir, las frecuencias relativas condicionales de los valores de Y_1, Y_2, Y_3 son iguales en X_1 y X_2. Diremos que las variables son independientes pues las frecuencias relativas condicionales de una de ellas no dependen del valor de la otra.
- Los casos a) y c) tienen solo interés teórico pues lo más frecuente es encontrarse en el caso b) donde hay una asociación parcial en los datos, que sería el caso correspondiente al ejemplo, En este caso las frecuencias absolutas dobles de los valores de Y_1, Y_2, Y_3 no son proporcionales en X_1 y X_2. Es decir, las frecuencias relativas condicionales de los valores de Y_1, Y_2, Y_3 son diferentes en X_1 y X_2. Lo que nos va interesar desde el punto de vista estadístico es saber cuando esta relación seria estadísticamente significativa y cuál sería la intensidad de esta relación.

5. Independencia

Como hemos visto, un método para estudiar la posible asociación o independencia entre dos variables es mediante el estudio de las distribuciones condicionales. Pero hay otras propiedades que nos pueden informar de la independencia (Batanero y Díaz, 2008). Veámoslo con un ejemplo:

Ejemplo 2. Práctica deportiva. Un investigador quiere estudiar si hay asociación entre la práctica deportiva y la sensación de bienestar. Extrae una muestra aleatoria de 250 sujetos. Los datos aparecen a continuación en la Tabla 4.

Sensación de bienestar	Práctica deportiva		Total
	Sí	No	
sí	90	60	150
no	60	40	100
Total	150	100	250

Tabla 4. Práctica deportiva

Al proponer este ejercicio a los estudiantes, muchos podrían pensar que las variables están relacionadas, pues la celda donde hay mayor frecuencia es en las personas que tienen sensación de bienestar y practica deporte. ¡Pero sería un error basar el juicio de asociación en los datos de una sola celda de la tabla! Para analizar la asociación podemos comparar la proporción de personas con sensación de bienestar entre los que practican deporte y los que no. Obtendríamos la tabla de frecuencias relativas condicionales por columnas (Tabla 5), donde observamos que la proporción de personas con y sin sensación de bienestar es la misma entre los que practican deporte y los que no. Es decir frecuencias relativas condicionales por columnas son iguales. Por otro lado, estas frecuencias relativas condicionales por columnas son también iguales a las frecuencias relativas marginales por filas, es decir a la proporción de personas con y sin insomnio en el total de la muestra o lo que es lo mismo, la distribución de *X* no cambia cuando se condiciona por un valor de *Y*.

Sensación de bienestar	Práctica deportiva		Total
	Sí	No	
sí	90/150=0,6	60/100=0,6	150/250=0,6
no	60/150=0,4	40/100=0,4	100/250=0,4
Total	150	100	250

Tabla 5. Frecuencias relativas condicionales por columnas

En consecuencia, una primera propiedad es que la variable *X* es independiente de *Y* si todas las distribuciones de frecuencias relativas que se obtienen al condicionar *X* por diferentes valores de $Y= y_j$ son iguales entre si e iguales a la distribución marginal de la variable *X*, es decir, cuando se verifica para todo par de valores i, j.

$$h(x_i|y_j)= h_{i.}$$

En el caso de independencia, se cumplen, además, las propiedades siguientes, que podemos comprobar en el ejemplo:

- La frecuencia relativa marginal de cualquier valor de *Y* condicionada por un valor de *X*, es igual para todos los valores de *X*, es decir *Y* no depende de *X*: $h(y_j|x_i)= h_{.j}$,
- La frecuencia relativa doble es igual al producto de las frecuencias relativas marginales de su fila y su columna: $h_{ij}= h_{i.}\ h_{.j}$, para todo i, j.

Esta última propiedad nos da un método de cálculo de las frecuencias teóricas en caso de independencia.

Para ello, desarrollamos la fórmula anterior: $h_{ij}= h_{i.}\ h_{.j}$; $\frac{f_{ij}}{n}=\frac{f_{i.}}{n}\times\frac{f_{.j}}{n}$ y simplificando, obtenemos: $f_{ij}=\frac{f_{i.}\times f_{.j}}{n}$

El valor $e_{ij} = \frac{f_{i.} \times f_{.j}}{n}$ se denomina frecuencias esperadas en caso de independencia; calculadas para el ejemplo tendremos la Tabla 6, donde podemos comprobar que, para este caso, las frecuencias observadas son iguales a las frecuencias esperadas, y por tanto corresponde al caso de independencia.

Sensación de bienestar	Práctica deportiva	
	sí (y_1)	no (y_2)
sí (x_1)	90 (e_{11})	60 (e_{12})
no (x_2)	60 (e_{21})	40 (e_{22})

Tabla 6. Frecuencias esperadas en caso de independencia

6. Signo de la asociación en tablas 2x2

En las tablas 2x2 podemos diferenciar entre dependencia directa y dependencia inversa y algunas celdas nos informan del signo de la asociación en este caso, como podemos comprobar en el siguiente ejemplo.

Ejemplo 3. Alergia (Estepa, 1993). Se quiere saber si sufrir o no de alergia tiene relación con llevar una vida sedentaria (llevar una vida sin realizar ningún tipo de ejercicio físico). Los datos de 300 sujetos se presentan en la Tabla 7, donde observamos que hay 130 personas con alergia y vida sedentaria, es decir con los dos caracteres al mismo tiempo y asimismo, 120 personas sin alergia y con vida no sedentaria. En total 250 de las 300 personas de la muestra o tienen a la vez los dos caracteres o no tienen ninguno. Estas dos celdas (presencia-presencia y ausencia-ausencia) informan que la asociación en la tabla es directa (Tabla 8).

Forma de vida	Sufre alergia	No sufre alergia
Sedentaria	130	30
No sedentaria	20	120

Tabla 7. Frecuencias absolutas dobles

Forma de vida	Sufre alergia	No sufre alergia
Sedentaria	Dep. directa	Dep. inversa
No sedentaria	Dep. inversa	Dep. directa

Tabla 8. Significados de las celdas en una tabla 2x2

Por el contrario en las otras dos celdas, se da un solo carácter y el otro no y serían las celdas favorables a una asociación inversa. En el ejemplo hay solo 50 casos en estas dos celdas y deducimos, en consecuencia que la asociación es directa.

7. Medidas de asociación en tablas 2x2

En el ejemplo anterior rechazamos la independencia de las variables, pues no se cumplen ninguna de las propiedades vistas para el caso de independencia. Para poder juzgar si la asociación es alta o baja, el siguiente paso sería calcular algún valor que mida la intensidad de la asociación. A continuación vamos a mostrar algunos coeficientes que sirven para medir esta intensidad, así como el signo de la asociación y tienen una interpretación sencilla (Ato y López, 1996):

- Si el coeficiente es positivo la asociación es directa, es decir, *A* y *B* suelen suceder juntos. Por tanto, si se da *A* suele darse *B*; por tanto habrá muchos

casos en la celda f_{11}. Por otro lado, si no se da *A*, lo más frecuente es que tampoco se de B, por tanto habrá muchos casos en la celda f_{22}. Por ejemplo "ser rubio" y "ojos claros" tendría asociación directa, pues habrá muchos casos de rubios con ojos claros y también de morenos con ojos oscuros.

- Si el coeficiente es negativo la asociación es inversa, es decir, si se da *A* no suele ocurrir *B* y si se da *B* no suele ocurrir *A*. Habría mayor frecuencia en las celdas f_{21} y f_{12}. Por ejemplo, "hacer deporte habitualmente" tendría una asociación negativa con la variable "estar obeso".
- Si el coeficiente es nulo no existe asociación, es decir, son independientes. No se encuentra un patrón en las diferentes celdas.

A continuación calculamos e interpretamos algunos de estos coeficientes, usando el siguiente ejemplo:

Ejemplo 4. Placebo. Un grupo de 230 personas aquejadas de insomnio fue dividido aleatoriamente en dos subgrupos. Al primer grupo se ofreció unas píldoras realmente somníferas para que tomaran una cada noche antes de acostarse, y al otro se ofreció un placebo (médicamente sin efecto somnífero). Al cabo de un mes fueron interrogados sobre la eficacia de las pastillas tomadas, con el resultado mostrado en la tabla 9:

	Dicen ser eficaces	Dicen no ser eficaces	Total
Píldoras somníferas	**80**	**15**	**95**
Placebo	**56**	**79**	**135**
Total	**136**	**94**	**230**

Tabla 9. Personas aquejadas de insomnio

7.1. Riesgo relativo

Este coeficiente se puede calcular por filas y columnas. El riesgo relativo por columnas indica cuanto más probable es la presencia de *A* cuando se da *B* que entre aquellos casos en los que no se da *B*, y se calcula mediante la siguiente fórmula:

$$RR_{columnas} = \frac{P(A/B)}{P(A/noB)} = \frac{f_{11}/f_{.1}}{f_{12}/f_{.2}} = \frac{f_{11}f_{.2}}{f_{.1}f_{12}}$$

En el ejemplo tendríamos:

$$RR_{columnas} = \frac{P(Eficaces/somnífero)}{P(Eficaces/Placebo)} = \frac{80x94}{136x15} = 3{,}68$$

El $RR_{columnas} > 1$, nos dice que es 3,68 veces más probable sentir un efecto con el somnífero que con el placebo. Por tanto indica una asociación positiva y fuerte entre el medicamento y el efecto.

El riesgo relativo por filas indica cuanto más probable es la presencia de *B* con *A* que entre aquellos que no pose en *A* y en nuestro ejemplo se obtendría el siguiente valor:

$$RR_{filas} = \frac{P(somnífero/Eficaces)}{P(Placebo/NoEficaces)} = \frac{80x135}{56x95} = \frac{10800}{5320} = 2{,}03$$

El $RR_{filas} > 1$, nos dice que existe asociación positiva y que es 2,03 veces más fácil haber tomado píldoras somníferas entre las personas que consiguieron el efecto somnífero y por tanto las consideraron eficaces que entre las que no consideraron eficaces las píldoras tomadas.

Como vemos en el ejemplo, los riesgos relativos por filas y columnas pueden no coincidir. Esto es debido a que estamos considerando una de las dos variables como dependiente de la otra; por tanto la fórmula varía dependiendo de cuál sea la variable que se considere independiente (es una medida de asociación asimétrica). En general, obtenemos la siguiente interpretación

- El $RR = 1$, informa que no hay asociación entre las variables.
- El $RR > 1$, nos dice que existe asociación positiva. Es lo que ocurre en el ejemplo mostrado.
- El $RR < 1$, indica que existe una asociación negativa.

7.2. Razón de productos cruzados

Este coeficiente, como su nombre indica, es el cociente entre el producto de las celdas favorables a la asociación positiva y las favorables a la asociación negativa y se calcula mediante la siguiente fórmula:

$$RC = \frac{f_{11} f_{22}}{f_{12} f_{21}} = \frac{f_{11} / f_{21}}{f_{12} / f_{22}} = \frac{C_1}{C_2}$$

Como vemos, la razón de productos cruzados es razón entre dos cocientes: C_1 es la razón de casos en que se presenta *A* y los que no se presenta *A* cuando está presente *B*. C_2 es la razón de casos *A* y no *A* cuando no está presente el factor *B*. En el ejemplo anterior tenemos la siguiente razón de productos cruzados:

$$RC = \frac{f_{11} f_{22}}{f_{12} f_{21}} = \frac{80x79}{15x56} = \frac{6320}{840} = 7{,}52$$

La interpretación de este valor es que entre los pacientes que encontraron eficaz el medicamento, hay 80 que tomaron píldoras somníferas por cada 56 que tomaron placebo. Por otro lado, entre los pacientes que no sintieron efecto hay 15 pacientes con píldoras somníferas por cada 79 con placebo. Se ve en el ejemplo que el somnífero fue mucho más eficaz que el placebo.

Conviene observar que la Razón de productos cruzados es también una medida asimétrica, es decir, *A* es la variable dependiente y *B* la independiente y si cambiamos la variable dependiente e independiente cambiará su valor. Se interpreta en la siguiente manera:

- El $RC = 1$, implica que las razones entre los casos en que aparecen *A* y no *A*, cuando está *B*, y los casos de *A* y no *A* cuando no está presente *B*, son iguales. Las variables serían independientes.
- El $RC < 1$, implica que la razón de casos en que aparece *A* y no *A*, cuando está presente *B* es menor que la de casos de *A* y no *A* cuando no está presente *B*. Hay una asociación inversa.
- De forma similar se interpreta cuando $RC > 1$ que implica una asociación directa.

7.3. Coeficiente Phi de Pearson

Las medidas anteriores son muy intuitivas y no requieren de procedimientos de inferencia, de modo que su comprensión está al alcance de los estudiantes de secundaria. Para el último curso de Bachillerato, donde los estudiantes de ciencias sociales han estudiado el contraste de hipótesis, podríamos también introducirles, en primer lugar, el coeficiente Phi de Pearson. Dicho coeficiente está basado en el valor Chi-cuadrado, el cual trata de calcular la distancia entre las frecuencias observadas y las frecuencias esperadas en caso de independencia. Se obtiene el valor 0 en caso de que las frecuencias observadas sean todas iguales a las esperadas, que, como vimos, ocurre sólo si las variables son independientes

$$\chi^2_{\exp} = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

.

Al ser este coeficiente una suma, cuantos más sumandos tengamos, mayor es el valor que toma, esto quiere decir que al aumentar el número de filas o columnas aparecerán más sumandos. Por otro lado, al depender de las frecuencias observadas y esperadas, es decir, será mayor al aumentar el valor n del tamaño de la muestra; es decir, el coeficiente depende del tamaño de la muestra. Para conseguir un coeficiente que no dependa del tamaño de la muestra se define el coeficiente Phi, que toma valores entre -1 y 1: de la forma siguiente:

$$\Phi = \sqrt{\chi^2 / n}$$

En el caso de tablas 2x2, se puede demostrar fácilmente que la formula quedaría de la siguiente forma:

$$\Phi = \frac{f_{11} f_{22} - f_{12} f_{21}}{\sqrt{f_1 \cdot f_{\cdot 1} f_2 \cdot f_{\cdot 2}}}$$

- Cuando la dependencia es directa y perfecta, todos los casos están en las celdas f_{11} y f_{22}, por tanto el coeficiente toma el valor 1. En general, si el coeficiente es positivo, la dependencia es directa y más alta cuanto más se acerque a 1.
- Cuando la dependencia es inversa y perfecta, todos los casos están en las celdas f_{12} y f_{21} y por tanto se obtiene un valor (-1). Si el coeficiente es negativo, la dependencia es inversa y más alta cuanto más se acerque a -1.
- El valor 0 se obtiene cuando hay independencia.

Para aplicar este coeficiente al ejemplo 4, comenzamos calculando las frecuencias esperadas (Tabla 10).

	Dicen ser eficaces	Dicen no ser eficaces
Píldoras somníferas	**(95x136)/230=56,17**	**(95x94)/230=38,83**
Placebo	**(135x136)/230=79,83**	**(135x94)/230=55,17**

Tabla 10. Frecuencias esperadas

A partir de ellas obtenemos el valor Chi-cuadrado, que es distinto de 0 y por tanto, los datos no son independientes. Para una evaluación más exacta calculamos el valor Phi de Pearson:

$$\chi^2_{\exp} = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(80-56{,}17)^2}{56{,}174} + \frac{(15-38{,}83)^2}{38{,}83} + \frac{(56-79{,}83)^2}{79{,}826} + \frac{(79-55{,}17)^2}{55{,}17} = 42{,}13$$

$$\Phi = \sqrt{\chi^2 / n} = \sqrt{42{,}13 / 230} = 0{,}44$$

Observamos que el valor es positivo (dependencia directa, moderada) y, en efecto, aparecen más datos en la diagonal principal f_{11} y f_{22} que en la otra diagonal.

8. Conclusiones

Los ejemplos y coeficientes presentados están al alcance de la comprensión de los estudiantes de los últimos cursos de secundaria excepto quizás el coeficiente Phi, pero este podría ser fácilmente comprensible por los estudiantes de Bachillerato, especialmente en el caso de los de ciencias sociales en el segundo curso, puesto que, una vez estudiado el contraste de hipótesis podría también enseñárseles el contraste Chi- cuadrado.

Pensamos que el resto de las estrategias mostradas (como comparar frecuencias condicionales entre sí o con las marginales, comparar frecuencias observadas con las esperadas en caso de independencia o bien el cálculo de los riesgos relativos y razón de productos cruzados son recursos muy interesantes para evitar a los alumnos los sesgos de razonamiento sobre las tablas 2x2 que se describieron en los antecedentes.

Actualmente hay a disposición de los ciudadanos (por ejemplo en Internet) toda clase de datos, lo que requiere la necesidad de desarrollar una mejor comunicación entre los productores de estadísticas y los consumidores. Muchas de dichas informaciones son presentadas en forma de tablas de contingencia, por lo que una persona estadísticamente culta debiera ser capaz de comprender e interpretar este objeto matemático, que es considerado como sencillo, y consideramos que no recibe la importancia que merece. El currículo escolar ha ofrecido hasta ahora pocas posibilidades para trabajar con este tipo de datos, pues la enseñanza de la estadística usualmente se reduce al estudio de variables cuantitativas. Esperamos que nuestro trabajo contribuya a soslayar este olvido.

Agradecimientos: Proyecto EDU2010-14947 y becas FPU AP2009-2807 y FPI BES-2011-044684 (MCINN-FEDER) y Grupo FQM16 (Junta de Andalucía).

Bibliografía

Alloy, L.B. y Tabachnik, N. (1984). *Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information*. *Psychological Review, 91,* 112-149.

Ato, M. y López, J.J. (1996). *Análisis estadístico para datos categóricos.* Sintesis Psicología, Madrid.

Batanero, C. y Díaz, C. (2008). *Análisis de datos con Statgraphics.* Departamento de Didáctica de la Matemática, Granada.

Cazorla, I. (2002). *A relaçao entre a habilidades viso-pictóricas e o dominio de conceitos estatísticos na leitura de gráficos*. Tesis Doctoral. Universidad de Campinas.

Estepa, A. (1993). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores*. Tesis Doctoral. Universidad de Granada.

Inhelder, B. y Piaget, J. (1955). *De la logique de l´enfant à la logique de l´adolescent*. Presses Universitaires de France, París.

Jenkins, H. M. y Ward, W.C. (1965). *Judgment of the contingency between responses and outcomes*, *Psychological Monographs,* 79, 1-17.

Jennings, D. L., Amabile, T. M. y Ross, L. (1982). *Informal covariation assessment: Data-based versus theory-based judgments*. En: D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases,* 211-230. Cambridge University Press, Nueva York.

Postigo, Y. y Pozo, J.I. (2000). *Cuando una gráfica vale más que 1000 datos: la interpretación de gráficas por alumnos adolescentes. Revista Infancia y Aprendizaje*, 90, 89 - 110.

Wild, C., y Pfannkuch, M. (1999). *Statistical thinking in empirical enquiry (con discusión). International Statistical Review*, 67(3), 223-265.

Wright, J. C. y Murphy, G.L. (1984). *The utility of theories in intuitive statistics: the robustness of theory-based judgments*. *Journal of Experimental Psychology General,* 113(2), 301-322.

Gustavo R. Cañadas. Lic. en c.c. y t.t. Estadísticas (Universidad de Granada), Master en Metodología (UNED), Master y Doctorado en Didáctica de la Matemática (Universidad de Granada). Fue becado en el Plan de Formación del Profesorado Universitario (2009). Ha publicado trabajos relacionados con las tablas de contingencia.
Email: **grcanadas@ugr.es**

José M. Contreras García. Profesor ayudante doctor de la universidad de Granada. Lic. en ciencias matemáticas, lic. en c.c. y t.t. estadísticas, diploma de estudios avanzados en Estadística e I.O., máster en didáctica de la matemática, máster en estadística aplicada y doctor en didáctica de la matemática. Publicaciones en didáctica de la probabilidad. **jmcontreras@ugr.es**

Pedro Arteaga. Lic. en Matemáticas (Universidad Complutense), Master y doctorado europeo en Didáctica de las Matemáticas (Universidad de Granada). Fue becado en el Plan de Formación del Profesorado Universitario (2007). Ha publicado trabajos relacionados con la comprensión de gráficos y el trabajo con proyectos estadísticos

María Magdalena Gea. Lic. en Matemáticas, Lic. en Ciencias y Técnicas Estadísticas, Máster en Estadística Aplicada y el Diploma de estudios avanzados. Su investigación se desarrolla en torno a la enseñanza y aprendizaje de la asociación estadística (correlación y regresión) en el marco del Plan de Formación de Personal Investigador (2011)