Chi cuadrado de Pearson

para dos variables nominales

Vicente Manzano Arrondo – 2014

Situación de partida

1	1	1
2	1	1
3	1	2
4	1	1
5	1	2
6	1	1
7	1	1
8	1	1
9	1	2
10	1	1
11	1	1
12	1	2
13	1	2
14	1	1
15	1	1
16	1	2
17	1	1
18	1	3
19	1	2
20	1	2
21	1	3
22	1	2
23	1	1
24	1	1
25	2	3
26	2	1
27	2	3
28	2	3
29	2	1
30	2	2
31	2	3
32	2	2
33	2	3
34	2	3
35	2	3
36	2	3
37	2	3
38	2	3
39	2	3
40	2	2

Queremos saber si los estudiantes de psicología y de económicas difieren en sus gustos literarios. Para comprobarlo, escogemos una muestra aleatoria de 40 estudiantes entre estas dos titulaciones y les presentamos tres títulos, cada uno de ellos representativos de un tipo de literatura. Pongamos A:“El paisaje de Roubeau”, B:“Alguien te está esperando” y C:“Las aventuras de Oliver Kelvin”. El listado de las respuestas figura a la izquierda de este texto.

Cada fila representa a una persona entrevistada. La primera columna es el código numérico de la persona entrevistada. La segunda se refiere a la titulación que se cursa (1: psicología; 2: económicas). Y la tercera columna indica el libro escogido (1:A; 2:B; 3:C).

Estudio de la relación. Tabla de contingencia

Para estudiar la relación, lo primero que podemos llevar a cabo es una tabla de frecuencias de cada variable por separado, lo que genera:

Esta información es útil para tener una idea de cómo se distribuyen ambas variables, pero no nos dice nada sobre su relación. Que lo estén implicaría que estudiantes de psicología mostrarían frecuencias de predilección diferentes a las de los estudiantes de economía. Para verlo, necesitamos que las tablas se realicen para cada titulación. Por ejemplo:

Psicología

Xi	fi
1	13
2	9
3	2
Total	24

Economía

Xi	fi
1	2
2	3
3	11
Total	16

fi

Xi	Psi.	Eco.
1	13	2
2	9	3
3	2	11
Total	24	16

Ya puestos, he generado tres tablas. Las dos primeras son tablas de frecuencia habituales, una para psicología y otra para economía. La tercera tabla, como puedes observar, es una fusión de las dos anteriores. Como comparten los mismos valores para los libros leídos (códigos 1, 2 y 3), podemos fundir ambas tablas, reproduciendo solo las frecuencias. Ese procedimiento nos permite comparar con más facilidad la predilecciones de los estudiantes de ambas titulaciones. En esta tercera tabla contamos además con el total por titulación, aunque no con el total por libro. Vamos a

añadir esta última información y, para aprovechar mejor el espacio, pongamos la tabla más horizontal:

	Frecuencias observadas	Libro A	Libro B	Libro C	Total
Titulación	Psicología	13	9	2	24
	Economía	2	3	11	16
	Total	15	12	13	40

Lo que ves recibe el nombre de “Tabla de contingencia”. A diferencia de una tabla de frecuencias, que maneja una sola dimensión, en la tabla de contingencia manejamos dos. La combinación de filas por columnas genera lo que llamamos *celdas* o *casillas*. En ellas podemos ver, por ejemplo, que hay tres estudiantes de economía que han escogido el libro B, o 13 de psicología que prefieren el título A. Las sumas de las frecuencias de las casillas, sea por filas o por columnas, se denominan *puntuaciones marginales*, lo que indica literalmente que se encuentran en los márgenes de la tabla. Las puntuaciones marginales son como tablas de frecuencia unidimensionales. Como puedes observar, las puntuaciones marginales de las filas son como la tabla de frecuencias de la variable “Titulación”, mientras que las marginales de columnas coinciden con la tabla de frecuencias de la variable “Libro”.

Observar la tabla de contingencia nos permite obtener conclusiones interesantes al objetivo del estudio. Los estudiantes de psicología prefieren A, seguido de B y, por último C. Mientras que en el caso de los estudiantes de economía ocurre lo contrario. La mayoría de psicología prefiere A. La mayoría de economía, C. A la vista de esta tabla, podríamos concluir, al menos respecto a nuestra muestra de 40 estudiantes, que hay una clara relación entre la titulación que se cursa y el estilo de literatura que gusta.

Una estrategia para tener claro que existe relación entre dos variables que hemos dispuesto en una tabla de contingencia es responder a preguntas relativas a una de las dos variables. Si la respuesta es “depende del valor de la otra”, entonces hay relación. Por ejemplo, “Los estudiantes, ¿qué tipo de literatura prefieren, A, B o C?”. Respuesta: “Depende, pues aunque en términos generales parece que se prefiere A, después C y, por último B (pero con poca diferencia), ocurre que los estudiantes de psicología prefieren claramente A, mientras que en economía se prefiere claramente C”. O bien “Estoy pensando en vender libros con un determinado estilo en alguna titulación, dime ¿hay más estudiantes de economía o de psicología?”. Respuesta “Depende, pues aunque en términos generales hay más estudiantes de psicología que de economía, esto varía mucho dentro de los gustos literarios; así, en el estilo A hay claramente más de psicología, pero en el estilo C la mayoría son de economía”.

Cuantificación. Chi cuadrado de Pearson

La tabla de contingencia arroja mucha luz a nuestro estudio, pero no basta con interpretar la tabla. Buscamos conseguir una expresión numérica que indique el grado en que existe relación. En términos generales, una buena estrategia para cuantificar una relación es idear un índice o estadístico que mida la distancia que existe entre lo que ocurre y lo que cabría ocurrir si no hubiera absolutamente nada de relación, es decir, si ambas variables fueran totalmente independientes. Si no hay ninguna distancia entre ambas situaciones, el índice suministra el valor 0. Conforme más lejos se encuentre de 0, estará indicando mayor grado de relación.

Para poner eso en práctica en el caso de relación de dos variables nominales expresada mediante una tabla de contingencia, necesitamos identificar qué ocurriría en la tabla si no existiera relación. Dado que nos importa la relación entre ambas variables y no cada una de ellas por separado, los marginales de la tabla permanecen del mismo modo. En otras palabras: a la relación le da lo mismo que haya más o menos estudiantes de una u otra titulación o que unos u otros libros se prefieran más. Lo que importa es "dado un total de libros y estudiantes ¿cómo se relacionan entre ellos?". Así que partimos de la tabla siguiente, con el objetivo de deducir qué debería ocurrir en el interior de las celdas o casillas para concluir que no existe relación alguna:

		Libro			
		A	B	C	Total
Titulación	Psicología				24
	Economía				16
	Total	15	12	13	40

Si no existiera ninguna relación, ante por ejemplo la pregunta "¿Hay más gente de economía o de psicología?" No responderíamos "Depende de en qué grupo de preferencia de lectura nos encontremos". Si observas las puntuaciones marginales de la titulación, hay 24 estudiantes de psicología y 16 de economía, es decir, un 60% de psicología y un 40% de economía. Pues bien, si no existiera relación alguna, deberíamos observar exactamente lo mismo (60% y 40%) en cada uno de los tres grupos de preferencia literaria. En el caso del grupo que ha preferido el libro A, dado que el marginal es 15, hablamos entonces de 15*60/100 = 9 estudiantes de psicología y 15*40/100 = 6 estudiantes de economía. Si hacemos esto mismo con los otros dos libros, construimos una nueva tabla de contingencia que respeta los marginales pero que contiene en las casillas las *frecuencias esperadas* si no existiera ninguna relación.

Frecuencias esperadas		Libro			
		A	B	C	Total
Titulación	Psicología	9	7,2	7,8	24
	Economía	6	4,8	5,2	16
	Total	15	12	13	40

Hay que reconocer que no es posible observar algo 7,2 veces. Es un inconveniente en el cálculo de las frecuencias esperadas. Y es lo que hay.

Fíjate cómo hemos conseguido la frecuencia esperada de preferencias por el libro A de estudiantes de psicología, f_e = 9. Primero hemos dividido 24 (el total de la fila de psicología) entre 40 (total general), dando por resultado un 60%. Después hemos aplicado este porcentaje al total de elecciones del libro A (15). En definitiva, la operación ha consistido en aplicar 24*15/40 = 9. De forma más descriptiva:

Total fila *psicología*(24) x total columna *A*(15) / total(40) = casilla *psicología*(9)

Es una buena regla general: para obtener la frecuencia esperada de la celda de la fila F y la columna C, lo que hacemos es multiplicar los marginales de la fila F y la columna C y dividir el resultado entre el tamaño de la muestra o total de frecuencias. Observa:

fe	A	B	C	Total
Psicología	24*15/40=9	24*12/40=7,2	24*13/40=7,8	24
Economía	16*15/40=6	16*12/40=4,8	16*13/40=5,2	16
Total	15	12	13	40

Ya sabemos cómo calcular las frecuencias esperadas (f_e) a partir de las puntuaciones marginales e ignorando las frecuencias observadas (f_o). Ahora nos enfrentamos a otro problema: si tenemos 6 celdas o casillas y por tanto 6 diferencias f_o-f_e ¿cómo obtener un solo número que represente a las 6 diferencias? En efecto, podríamos sumar todas las celdas, $\Sigma(f_o-f_e)$. Pero esto tiene un inconveniente importante. Dado que finalmente todas las frecuencias, sean observadas o esperadas, han de sumar lo mismo, las diferencias por exceso se contrarrestan con las diferencias por defecto y, finalmente, esa suma siempre daría 0 como resultado. Así que idea descartada. Para comprobarlo, observa la siguiente tabla, donde figuran todas las diferencias f_o-f_e.

fo-fe	A	B	C	Total
Psicología	4	1,8	-5,8	0
Economía	-4	-1,8	5,8	0
Total	0	0	0	0

La solución aritmética, como podrías sospechar, es la misma a la que acudimos para calcular la varianza: elevar las diferencias al cuadrado. No obstante, la suma de las diferencias cuadráticas entre frecuencias, $\Sigma(f_o-f_e)^2$, tampoco es una idea perfecta. Así, por ejemplo, si tenemos muchos datos, sumamos más diferencias cuadráticas que si tuviéramos pocos. Y más datos no significa más discrepancia sino solo más datos y punto. Sería recomendable, como hicimos con la varianza, no quedarnos con la suma sino con la media. Pero esto no fue lo que se le ocurrió a Karl Pearson (1857 – 1936). Lo que hizo Pearson fue dividir cada diferencia cuadrática entre la frecuencia esperada ($[f_o-f_e]^2/f_e$). Esta idea permite expresar la distancia en la escala de cantidades que se está manejando. Si tenemos muchos datos y poco valores, por ejemplo, las frecuencias serán muy elevadas y las distancias menos relevantes. Este recurso está muy bien, aunque como vamos a ver no resuelve todos los problemas. Pearson utilizó la letra griega χ (en otras grafías: χ, X, χ...), que se lee *chi* o *ji*. Como las diferencias son cuadráticas, se la conoce como Chi cuadrado de Pearson y se simboliza con χ^2. Por lo tanto, la expresión de cálculo es:

$$\chi^2 = \sum \frac{(fo_i - fe_i)^2}{fe_i}$$

Para la primera celda o casilla, correspondiente a los estudiantes de psicología que prefieren el libro A, el contenido es $(13 - 9)^2 / 9 = 1,7777...$

La siguiente tabla contiene los componentes del cálculo: cada celda muestra el resultado de operar $(f_o-f_e)^2/f_e$. El total es precisamente el valor de la chi cuadrado (χ^2=16,351).

Componentes de la χ2 de Pearson		Libro A	Libro B	Libro C	Total
Titulación	Psicología	1,778	0,450	4,313	
	Economía	2,667	0,675	6,469	
	Total				16,351

Interpretación. V de Cramer

¿Qué significa χ^2=16,351? Desde luego, no es χ^2=0, situación en la que concluiríamos sin problemas con ausencia de relación. La chi cuadrado va desde 0 hasta un valor que varía según el número de datos y el número de celdas. Eso de no contar con un máximo fijo dificulta bastante la interpretación. No obstante, un suizo, llamado Harald Cramer (1893 – 1985), muy interesante en asuntos diversos del mundo de la estadística, estuvo razonando matemáticamente hasta llegar a la conclusión de que el valor máximo que puede tener el invento es $n(k\text{-}1)$, donde n es el número de datos y k es el número de valores o categorías de la variable que tiene menos valores. En nuestro caso, n = 40 y k = 2, por lo que el valor máximo que podríamos obtener aplicando Chi cuadrado en tales condiciones es 40(2-1)= 40. Cramer propuso un índice, llamado V de Cramer, para tranformar la Chi cuadrado de Pearson. La V consiste en dividir la chi entre su máximo, por lo que el resultado va de 0 (no hay nada de relación) a 1 (relación máxima). Dado que χ^2 está elevada al cuadrado, la propuesta concreta de Cramer es (de paso, calculamos ya nuestra V) aplicar también una raíz cuadrada:

$$V = \sqrt{\frac{\chi^2}{n\,(k-1)}} = \sqrt{\frac{16{,}351}{40\,(2-1)}} = 0{,}64$$

El problema ahora es qué hacer con esa V = 0,64, es decir, cómo concluir si existe o no relación. Esto es lo que ha venido llamándose problema del *tamaño de efecto*. Lo hemos abordado ya. No obstante, reproduzco la argumentación para este primer caso.

El tamaño de efecto viene a ser sinónimo de *grado o medida de relación acotada o estandarizada*. Para cada índice o estadístico de relación (como ocurre con la chi cuadrado), nos enfrentamos a la tarea de interpretarlo, por lo que ideamos una estrategia que suministre un valor acotado o estandarizado (como ocurre con la V) y ahí tenemos el tamaño del efecto. Tal y como está, está bien. Pero en muchas ocasiones necesitamos traducir el continuo del efecto en una dicotomía: “al final, dime, ¿hay o no relación?”. Responder a esta pregunta no es un asunto únicamente estadístico, tiene que ver con muchos aspectos, como por ejemplo las consecuencias de equivocarse al concluir que hay relación. Cuanto peores sean las consecuencias, mayor deberá ser la exigencia de cuantía en V para concluir que hay relación. Por otro lado, sabemos que existe un “ruido” constante en la naturaleza, es decir, valores de relación no nulo entre cualesquiera dos variables que se nos ocurran. Podemos tomar, por ejemplo, el tipo de oreja de una persona (en un sistema de tres categorías, por ejemplo). Extraemos al azar 60 zonas del planeta e identificamos cuál es el tipo de oreja mayoritario. Lo anotamos. Y ahora vemos también en cada una de esas 60 zonas cuántas veces ha sobrevolado una especie de pato que abunde en el mundo: ninguna, menos de la media mundial, más de la media mundial. En estos momentos no se me ocurre una tontería más grande. Pues bien, con esos datos podemos construir una tabla de contingencia de tres filas (tipos de oreja) y tres columnas (frecuencia de sobrevolado del pato). Aunque te parezca que no debería existir ninguna relación entre ambas variables, cuando realices la experiencia obtendrás una chi cuadrado superior a 0 y, por tanto, también una V de Cramer superior a 0. Esto siempre ocurre. A eso le estoy llamando “ruido de la naturaleza”.

Así que si combinamos las consecuencias de error, el ruido de la naturaleza y otros elementos (como el marco teórico que señala el significado especial de algunos valores de relación en algunas situaciones concretas), tenemos que considerar unos valores concretos de la V (o de cualquier tamaño de efecto) para concluir que sí hay relación o no la hay, es una aventura difícil e infructuosa. No obstante, autores como Jacob Cohen

(1923 – 1998), han dado muchas vueltas a este asunto y nos han suministrado alguna guía. Ya la conocemos::

- De 0 a 0,10, podemos decir que no hay efecto (el grado de relación es ridículo, despreciable o achacable al *ruido*).
- Desde 0,10 hasta 0,30, el efecto es pequeño.
- Desde 0,30 hasta 0,50, el efecto es mediano o moderado.
- Y desde 0,50 hasta 1,00, el efecto es grande.

Pues ahí lo tienes: una solución operativa para tomar una decisión sobre la cuantía de una relación, en las situaciones donde no sepas a qué cosa mejor agarrarte. En nuestro caso y dado que V = 0,64, podemos concluir que al menos en el conjunto de la muestra hay relación y grande además.

En definitiva, pues, para cuantificar una relación entre dos variables nominales, calculamos la chi cuadrado de Pearson y la transformamos según la V de Cramer, que nos permite obtener una cuantía comprendida entre 0 (ausencia absoluta de relación) y 1 (relación máxima). Para concluir si existe relación, pedimos a V una cuantía mínima de 0,10, a partir de la cual interpretamos si se trata de un efecto pequeño (V ≥ 0,1), mediano (V ≥ 0,3) o grande (V ≥ 0,5).

Un ejemplo

Vive	Qué hace
2	2
2	3
1	2
1	3
2	3
2	1
1	3
2	3
2	1
1	3
2	2
1	3
2	2
1	3
2	3
2	2
1	3
2	3
2	1
2	2
2	3

Hemos visitado un barrio de Sevilla, preguntando a la personas con quienes nos encontramos si viven o no en ese barrio (Variable VIVE. Valores 1.Sí; 2.No). Las hemos encontrado en situaciones diferentes (variable QUÉ HACE): están paseando (1), se encuentran de compras (2) o consumiendo en un bar o similar (3). Los resultados se encuentran en la tabla de datos que observas a la izquierda de este texto. Y planteamos la pregunta ¿existe relación entre lo que esa persona está haciendo cuando es entrevistada y dónde vive (dentro o fuera del barrio)?

Se trata de un par de variables nominales, con 2 y 3 categorías. Es un problema idóneo para ser resuelto a partir del cálculo de una chi cuadrado de Pearson.

En primer lugar, construyamos la tabla de contingencia:

Fobs		Actividad			
		1	2	3	Total
Vive en el barrio	1	0	1	6	7
	2	3	5	6	14
	Total	3	6	12	21

Puede observarse, por ejemplo, que hay 5 personas que no viven en el barrio y que las hemos entrevistado mientras se encontraban de compras.

A simple vista, la tabla muestra una aglomeración de valores de las personas que viven en el barrio y han sido entrevistadas en el bar, mientras que quienes vienen de fuera distribuyen más el tipo de actividad que realizan en el lugar. Vamos a realizar el siguiente

paso en la consecución de la chi cuadrado: calcular las frecuencias esperadas. Recuerda: frecuencias *esperadas* si no existiera nada de relación entre ambas variables (dos variables totalmente independientes). En otras palabras: si conocer una variable no fuera en absoluto útil para saber algo de la otra, entonces qué frecuencias cabría esperar. Respuesta: las frecuencias esperadas.

Fesp		Actividad			
		1	2	3	Total
Vive en el barrio	1	1	2	4	7
	2	2	4	8	14
	Total	3	6	12	21

Por ejemplo: $fe_{12} = 7 \cdot 6 / 21 = 2$
Para calcular la Chi cuadrado, operamos en cada casilla, encontrando:

(fo-fe)^2/fe		Actividad			
		1	2	3	Total
Vive en el barrio	1	1	0,5	1	2,5
	2	0,5	0,25	0,5	1,25
	Total	1,5	0,75	1,5	3,75

Por ejemplo: $(1 - 2)^2 / 2 = 0,5$
El resultado de sumar todas las casillas es el valor de Chi cuadrado:

$$X^2 = \sum \frac{(fo_i - fe_i)^2}{fe_i} = 3,75$$

Para interpretar esta cuantía, recurrimos a la V de Cramer:

$$V = \sqrt{\frac{X^2}{n\,(k-1)}} = \sqrt{\frac{3,75}{21\,(2-1)}} = 0,42$$

Como podemos observar, se trata de un efecto mediano y por tanto parece acertado concluir lo que hemos apuntado ya estudiando la tabla de contingencia: existe relación entre dónde vive la persona entrevistada (dentro o fuera del barrio) y el tipo de actividad que realiza.

No olvidemos que nos estamos refiriendo a unos datos concretos. Esta conclusión es válida para las 21 personas entrevistadas. No podemos ir más allá. Para aspirar a concluir a nivel de la población en su conjunto, necesitamos poner en marcha otros procedimientos más que los aplicados aquí: el conjunto de las 21 personas debería provenir de un muestreo aleatorio y el valor de chi cuadrado se sometería a una prueba estadística para concluir si la relación puede o no defenderse a nivel de la población. Pero esto es otra historia, que abordaremos más adelante. Ahora, lo que tenemos a mano (y es mucho, lo más importante) es la descripción de la relación mediante la tabla de contingencia y el tamaño del efecto mediante la cuantificación Chi cuadrado, transformada con V de Cramer.