PRUEBA DE CHI CUADRADO (χ^2) PARA UNA SOLA MUESTRA

La prueba del χ^2 se usa para variables de distintos niveles de medición, incluyendo las de menor nivel, que son las nominales. Sirve para determinar si los datos obtenidos de una sola muestra presentan variaciones estadísticamente significativas respecto de la hipótesis nula.

Cuando formulamos una hipótesis de trabajo, simultáneamente definimos la hipótesis nula, que niega nuestra hipótesis de trabajo. De acuerdo a la hipótesis nula (H_0) las variaciones en la variable independiente no tienen correspondencia con las variaciones que pudiere haber de la variable dependiente. Es decir que existe "independencia estadística".[1] Las variaciones que pudiese encontrarse se deberían a factores aleatorios, ajenos a la variable independiente.

Para comprobar si esto es así (y, por lo tanto, deberíamos aceptar la H_0) o no (y, por ende, rechazarla), podemos someter los resultados obtenidos de nuestra muestra a una prueba de χ^2, que se postula con la siguiente ecuación:

$$\chi^2 = \frac{\Sigma\,(f_e - f_o)^2}{f_e}$$

Se trata de la razón entre la sumatoria de los cuadrados de las diferencias entre las frecuencias esperadas (f_e) y las frecuencias observadas (f_o) respecto de las frecuencias esperadas (f_e). Como toda razón, expresa una proporción; en este caso, la proporción entre las distancias observadas (elevadas al cuadrado) y las frecuencias esperadas.

Pero la aplicación del chi cuadrado no se puede hacer directamente. Es necesario, antes de ello, realizar dos pasos. Por una parte, establecer el nivel de significación (ά) con el cual vamos a trabajar, y determinar los grados de libertad de nuestra muestra.

El nivel de significación es arbitrario y se fija de antemano (usualmente entre 0.01 y 0.10, siendo el más usado el de 0.05). Los grados de libertad se establecen en función de la cantidad de celdas que tenemos, producto de las categorías de una variable o bien de la cantidad resultante del cruce de dos variables.

[1] Véase Unidad 3.

GRADOS DE LIBERTAD

Esta noción se refiere a la posibilidad que se tiene de establecer, en una distribución dada, valores arbitrarios sin modificar el marginal de dicha distribución. Así, en una variable con cinco categorías, podré establecer cuatro valores de manera arbitraria, ya que el quinto quedará determinado por la diferencia entre la sumatoria de los cuatro que establezco, y el marginal. Cuando tengo una variable, la fórmula para calcular los grados de libertad es

$$df = k - 1$$

siendo "k" el número de categorías que tengo.

PRUEBA DE CHI CUADRADO (χ^2) PARA MÁS DE UNA MUESTRA

Cuando trabajo con cuadros de doble entrada (dos variables), las categorías de la variable independiente constituyen, cada una, una muestra. Por ejemplo, si la variable independiente es sexo, tendré una muestra de hombres y otra de mujeres. En estos casos la forma de establecer los grados de libertad es

$$df = (c - 1) \,.\, (f - 1)$$

siendo "c" el número de columnas y "f" el número de filas.

Es decir que es el producto del número de celdas menos uno, por el número de filas menos uno.

CÁLCULO DE LAS FRECUENCIAS ESPERADAS

Las frecuencias esperadas (f_e) vienen dadas por la hipótesis nula (H_0), pero no siempre se puede establecer de manera inmediata. Esto solo es posible cuando trabajamos con

una variable, pero cuando tenemos cuadros de doble entrada la forma de establecer el valor de la frecuencia esperada de cada celda es el siguiente:

	Categoría 1	Categoría 2	Categoría 3	Marginal 1/2/3
Categoría A	*a*	*b*	*c*	(*a*+*b*+*c*)
Categoría B	*d*	*e*	*f*	(*d*+*e*+*f*)
Categoría C	*g*	*h*	*i*	(*g*+*h*+*i*)
Marginal A/B/C	(*a*+*d*+*g*)	(*b*+*e*+*h*)	(*c*+*f*+*i*)	N

Cálculo de la frecuencia esperada (f_e) para la celda *a*

$$\frac{(a+d+g)\,(a+b+c)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *b*

$$\frac{(b+e+h)\,(a+b+c)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *c*

$$\frac{(c+f+i)\,(a+b+c)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *d*

$$\frac{(a+d+g)\,(d+e+f)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *e*

$$\frac{(b+e+h)\,(d+e+f)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *f*

$$\frac{(c+f+i)\,(d+e+f)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *g*

$$\frac{(a+d+g)\,(g+h+i)}{}$$

N

Cálculo de la frecuencia esperada (f_e) para la celda *h*

$$\frac{(b+e+h)\,(g+h+i)}{N}$$

Cálculo de la frecuencia esperada (f_e) para la celda *i*

$$\frac{(c+f+i)\,(g+h+i)}{N}$$

Como puede observarse, el procedimiento es bien sencillo. Se trata de la razón entre el producto de los marginales de la celda considerada y el total (N).

COMPARACIÓN DEL VALOR OBTENIDO Y LECTURA DEL χ^2

Una vez que se obtiene el resultado de la ecuación, el número arrojado no tienen significación por sí mismo. En realidad lo obtenido es un parámetro para establecer la validez o no de mi hipótesis de trabajo. Si se observa la fórmula de χ^2

$$\chi^2 = \frac{\Sigma\,(f_e - f_o)^2}{f_e}$$

puede notarse que cuanto mayor es la diferencia entre las frecuencias observadas y las esperadas (f_o y f_e respectivamente), mayor será el numerador [$\Sigma\,(f_e - f_o)^2$] y, consecuentemente, también será mayor número que se obtenga. Una mayor diferencia indica, por otra parte, que es menos probable que las mismas se deban puramente al azar (que es lo que indicaría la H_o). Por esta razón, cuanto mayor sea el número obtenido, más probable es que podamos rechazar la hipótesis nula.

Decíamos que el número obtenido es simplemente un parámetro, es decir, un punto para comparar. ¿Y contra qué lo debemos comparar? Contra la tabla D, que es la distribución del χ^2. Para ello debemos considerar los grados de libertad (*df*) y el nivel de significación (ά) que hemos elegido. En los cabezales de las columnas de la tabla D

encontramos los niveles de significación, y en las filas, los grados de libertad. Cruzando ambos (columna y fila) llegamos a una celda con un número determinado.

Si el número que nosotros obtenemos mediante el cálculo de χ^2 es igual o mayor (= ó >) al que figura en la tabla, rechazamos la hipótesis nula (H_o) y validamos, en consecuencia, nuestra hipótesis de trabajo (H_1). Si, por el contrario, es inferior, debemos aceptar la hipótesis nula (H_o), quedando inválida nuestra hipótesis de trabajo (H_1).

			Lectura
Número obtenido de χ^2	>	Número de la tabla	Rechazo H_o. Acepto H_1
Número obtenido de χ^2	=	Número de la tabla	Rechazo H_o. Acepto H_1
Número obtenido de χ^2	<	Número de la tabla	Acepto H_o. Rechazo H_1

CÁLCULO DE χ^2

Vamos a ver prácticamente cómo se calcula el χ^2. Tomaremos dos ejemplos, para una y más de una muestra.

Ejemplo 1 (Cálculo para una muestra)

Suponemos que los compradores en los *shoppings* pertenecen a las clases altas de la sociedad. Para eso tomamos una muestra de 50 casos, de manera aleatoria, a quienes indagamos sobre su pertenencia social (para ello debemos tener un instrumento que nos permita inferir a qué clase social pertenecen). Los resultados nos arrojan lo siguiente:

Clase	Baja	Media–Baja	Media	Media–Alta	Alta
f_o	8	9	10	11	12

Para saber si el resultado obtenido es estadísticamente significativo sometemos esta muestra a una prueba de χ^2. La hipótesis nula queda formulada de la siguiente manera: "los compradores en los *shoppings* no pertenecen a una clase social específica", en razón de la cual, las frecuencias esperadas han de ser de 10 casos para cada celda (si la clase social no influye, no debe haber variación en una muestra tomada al azar, y si tal variación existe, ésta se debe a cuestiones contingentes, y no a una tendencia).

El nivel de significación que escogemos es ά = 0.05; procedemos, en consecuencia, a realizar el cálculo de χ^2.

Clase	Baja	Media–Baja	Media	Media–Alta	Alta
f_o	8	9	10	11	12
f_e	10	10	10	10	10
$f_e - f_o$	-2	-1	0	1	2
$(f_e - f_o)^2$	4	1	0	1	4

La sumatoria [Σ] de las diferencias al cuadrado [$(f_e - f_o)^2$] es = 4+1+0+1+4.

Reemplazando los términos obtenemos la siguiente ecuación:

$$\chi^2 = \frac{\Sigma (f_e - f_o)^2}{f_e} \qquad \chi^2 = \frac{10}{10} = 1$$

Estamos trabajando con 4 grados de libertad (k = 5–1), y el ά = 0.05; observamos en la tabla D el valor que corresponde a df = 4 y ά = 0.05 y el mismo es 9,488. Dado que 1 (el número obtenido) es inferior (<) al que figura en la tabla, aceptamos la hipótesis nula.

Número obtenido de χ^2	<	Número de la tabla	Acepto H_o. Rechazo H_1
1	<	9,488	Acepto H_o. Rechazo H_1

Tenemos que decir, en consecuencia, que los compradores en *shoppings* no pertenecen a una clase social específica, y que las variaciones que encontramos (8, 9, 10, 11, 12) se deben exclusivamente al azar.

Ejemplo 2 (Cálculo para cinco muestras)

Suponemos que la ideología política influye en la elección de los medios de prensa que se leen. En razón de ello, suponemos que la gente de derecha escoge *La Prensa*, que los de centro derecha leen *La Nación*, lo de centro leen *Clarín*, los de centro izquierda leen *Página/12* y los de izquierda *Le Monde Diplomatique*. Para ello construimos un instrumento que nos permite establecer la ideología, y tomamos una muestra al azar, obteniendo el siguiente resultado.

	Derecha	Centrode-recha	Centro	Centroiz-quierda	Izquierda	Total
La Prensa	34	12	8	5	2	61
La Nación	32	31	24	28	20	135
Clarín	15	55	68	61	34	233
Página/12	21	18	17	25	21	102
Le Monde	10	8	15	25	43	101
Total	112	124	132	144	120	632

Decidimos trabajar con un $\alpha = 0.05$. Con el procedimiento descrito anteriormente determinamos las frecuencias esperadas, de modo que nos quedan de la siguiente manera:

(Cuadro A)

	Derecha	Centrode-recha	Centro	Centroiz-quierda	Izquierda	Total
La Prensa	10,810126	11,968354	12,740506	13,898734	11,582278	61
La Nación	23,924050	26,487341	28,196202	30,759493	25,632911	135
Clarín	41,291139	45,715189	48,664557	53,088607	44,240506	233
Página/12	18,075949	20,012658	21,303797	23,240506	19,367088	102
Le Monde	17,898734	19,816455	21,094936	23,012658	19,177215	101
Total	112	124	132	144	120	632

La diferencia entre las f_o y las f_e es:

(Cuadro B)

-23,1898734	-0,03164557	4,74050633	8,89873418	9,58227848
-8,07594937	-4,51265823	4,19620253	2,75949367	5,63291139
26,2911392	-9,28481013	-19,335443	-7,91139241	10,2405063
-2,92405063	2,01265823	4,30379747	-1,75949367	-1,63291139
7,89873418	11,8164557	6,09493671	-1,98734177	-23,8227848

Los cuadrados de dichas diferencias son:

(Cuadro C)

537,770229	0,00100144	22,4724003	79,18747	91,8200609
65,2209582	20,3640843	17,6081157	7,61480532	31,7296908
691,224003	86,2076991	373,859357	62,5901298	104,86797
8,5500721	4,05079314	18,5226726	3,09581798	2,66639962
62,3900016	139,628625	37,1482535	3,94952732	567,525076

Las razones (divisiones) entre el cuadrado de la diferencia (Cuadro C) y la frecuencia esperada (Cuadro A) para cada caso son:

(Cuadro D)

49,746897	8,367E-05	1,7638546	5,6974592	7,927634
2,7261671	0,7688233	0,6244854	0,2475595	1,2378497
16,74025	1,8857562	7,6823746	1,1789748	2,3704062
0,4730082	0,2024116	0,8694541	0,1332079	0,1376768
3,4857215	7,046095	1,7610034	0,1716241	29,593717

La sumatoria de estos términos (todas las celdas del Cuadro D) es 144,472496. Para comparar con la tabla tenemos que calcular los grados de libertad (*df*). Para ello cuento las columnas y las filas que tiene el cuadro.

$$df = (c - 1)\,(f - 1) = (5 - 1)\,(5 - 1) = 4 \,.\, 4 = 16$$

En la tabla D observo que para $\acute{\alpha} = 0.05$ y $df = 16$ el valor que corresponde es 26,296. Es inferior al que me arrojó el cálculo de χ^2; por tal razón, debo rechazar la hipótesis nula (H_0). Se confirma así la hipótesis de trabajo (H_1).

COMENTARIOS ADICIONALES

Obsérvese que hemos utilizado variables ordinales (en el primer ejemplo) y nominales (en el segundo). En ambos casos el χ^2 nos ofrece, de igual modo, una respuesta acerca de la asociatividad de las mismas. ¿A qué se debe esto? A que esta prueba no nos indica si la asociación tiene algún sentido estipulado,[2] sino únicamente si existe o no asociación, dentro de los límites de seguridad fijados por nosotros mismos al establecer el nivel de significación. Con esto queremos decir que el orden en que se presenten los datos en las variables es indistinto, ya que al sumar todas las diferencias cuadráticas, eliminamos cualquier referencia a ese orden inicial. Por eso χ^2 es una prueba especialmente adecuada para las variables nominales, pese a que se la puede usar también con las ordinales.

Para comprobar esto vemos que si cambiamos el orden de las categorías del último ejemplo, en nada varía el resultado final, ya que las celdas cambiarán de ubicación,

[2] Véase "dirección de la asociación" en la Unidad 3.

pero los marginales serán los mismos, aunque en distinto orden. De modo tal, que el resultado final seguirá siendo el mismo.