Análisis de datos en los estudios epidemiológicos V
Prueba de Chi cuadrado y Análisis de la varianza

Julia García Salinero. Departamento de Investigación FUDEN.

Introducción

Continuamos el análisis de los estudios epidemiológicos, centrándonos en este capítulo en las pruebas estadísticas que debemos utilizar cuando queremos encontrar relación o asociación entre variables de estudio.

El capitulo anterior lo concluimos hablando de las pruebas de significación estadística. Como pudimos observar existen varias pruebas que nos permiten encontrar significación estadística entre las variables y para la elección de alguna de ellas teníamos que tener en cuenta entre otros aspectos: el tipo de variables que estamos estudiando y el número de categorías.

Dada la amplitud del tema, nos vamos a centrar en dos de las pruebas más utilizadas, que son, la Prueba Chi Cuadrado y el Análisis de la Varianza de una sola vía.

Prueba de la Chi cuadrado

Esta prueba de significación estadística nos permite encontrar relación o asociación entre dos variables de carácter cualitativo que se presentan únicamente según dos modalidades (dicotómicas).

Imaginemos que queremos realizar un estudio descriptivo en la población de Leganés para determinar si existe algún tipo de relación entre el hábito tabáquico y el nivel de estudios. Pues bien, realizado el correspondiente diseño metodológico, elegida una muestra representativa de esta población y obtenidos los datos, procedemos a su tabulación en tablas de contingencia y procedemos al cálculo de la Chi Cuadrado. Si al final de nuestro estudio concluimos que nuestras variables no están relacionadas decimos con un determinado nivel de confianza, previamente fijado, que éstas son independientes.

En uno de los capítulos dedicados a la estadística descriptiva hablamos de las distintas formas en que se podían ordenar los datos de una distribución de frecuencias, indicando entre ellas las tablas de contingencia. Una tabla de contingencia es una distribución con dos o más dimensiones (bidimensional) en la cual las frecuencias de dos o más variables se tabulan de manera cruzada. Las más frecuentemente utilizadas son las tablas de 2x2 (dos filas por dos columnas).

Julia García Salinero

Cálculo de la Chi cuadrado

El cálculo de la Chi cuadrado se realiza a partir de la fórmula:

$$\sum \frac{(f_e - f_t)^2}{f_t}$$

Vamos a proceder a su estudio a través de un ejemplo concreto. A un grupo de 100 pacientes que se quejaba de no dormir bien se les administró dos tipos de tratamientos distintos (A,B). Al concluir el estudio se les preguntó si habían dormido bien o mal y se obtuvieron los siguientes resultados:

Entre los 60 pacientes a los que se les había administrado el tratamiento A, 40 habían dormido bien y 20 habían dormido mal. Entre los 40 pacientes a los que se les había administrado el tratamiento B, 15 habían dormido bien y 25 mal. Pues bien, pretendemos saber si el tipo de tratamiento influye en dormir bien o mal.

Pasos a seguir:

En primer lugar presentamos los datos en una tabla de contingencia (2x2)

	Tratamiento A	Tratamiento B	
Durmieron bien	40 (33)	15 (22)	55
Durmieron mal	20 (27)	25 (18)	45
	60	40	100

1.- Planteamos la hipótesis nula (H0): no existen diferencias significativas entre los dos tratamientos, o bien, que las dos variables estudiadas son independientes

2.- Fijamos el nivel de confianza con el que queremos afirmar nuestro resultado. En nuestro caso fijamos el 95% , o lo que es lo mismo, asumimos un error del 5%, es decir una p< 0,05
3.- Calculamos las frecuencias teóricas, correspondientes a cada grupo, partiendo del supuesto de que ambas variables fuesen independientes y las colocamos entre paréntesis. Utilizamos la siguiente fórmula ft= total de filas x total de columnas / total global; Ejemplo ft =55 x 60/100= 33

4.- Calculamos la Chi cuadrado mediante la fórmula:

$$\sum \frac{(f_o - f_t)^2}{f_t}$$

$$\chi^2 = \frac{(40-33)^2}{33} + \frac{(15-22)^2}{22} + \frac{(20-27)^2}{27} + \frac{(25-18)^2}{18} = 8,249$$

5.- Calculamos los grados de libertad: Se obtiene mediante la formula: gl = (f-1) x (c-1).. gl = (2-1) x (2-1) = 1. En tablas de 2x2 los grados de libertad son siempre 1.

6.- Interpretación de los datos. Comparamos nuestros datos con el valor que se corresponde a esos grados de libertad y con ese nivel de confianza en la tabla. Si la Chi cuadrado obtenida es igual o menor al de las tablas, aceptamos la hipótesis nula o de independencia. En caso contrario la rechazamos y aceptamos la hipótesis alternativa H1 de que existe relación o dependencia entre estas variables.
En nuestro ejemplo procedemos a buscar en las tablas de la distribución de la Chi cuadrado el valor para 1 grado de libertad y una p < 0,05 y observamos que es de 3,841. Al ser nuestro valor de 8,249 > a 3, 841, rechazamos la hipótesis nula de no relación o independencia. Por tanto concluimos que con una probabilidad de error del 5% podemos afirmar que existe relación entre el tipo de tratamiento administrado y la calidad del sueño.

Análisis de la varianza

El procedimiento conocido como análisis de la varianza (ANOVA) es también una prueba de significación estadística muy utilizada. Se utiliza cuando queremos comparar una variable cualitativa con más de dos categorías, con una variable cuantitativa.

El análisis de la varianza se consigue aplicando la prueba F de Snedecor y comparando los valores que obtenemos en nuestro estudio con los valores correspondientes en dicha prueba a partir de un nivel de confianza fijado y con unos grados de libertad determinados en función de los factores en estudio.

F = VF / VR

Vamos a proceder a su estudio a través de un ejemplo concreto, como ya hicimos en la prueba anterior.

Julia García Salinero

Se realizó un estudio para ver si, en unas condiciones determinadas, la raza influye significativamente en la mayor o menor talla de las personas. Para ello se toman 8 individuos homogéneos con aquellas condiciones de raza blanca; 7 de raza negra y 6 de raza amarilla. Se miden sus estaturas (en centímetros) y se recogen en una tabla obteniéndose los siguientes valores:

Raza Blanca. Muestra 1	Raza negra. Muestra 2	Raza amarilla. Muestra 3
170	172	162
182	184	156
168	169	166
175	167	171
185	171	158
166	183	161
178	164	
180		

Dado que estamos intentando relacionar una variable cuantitativa (la talla) con otra cualitativa (la raza, categorizada en tres categorías), procederemos a realizar un análisis de la varianza.

Siendo V F = La varianza factorial o debida al factor de estudio.

V F = $\dfrac{\left(n_1 \cdot m_1^2 + n_2 m_2^2 + n_3 m_3^2\right) - N \cdot M^2}{k-1}$

Siendo V R = La varianza residual o debida al azar, es decir, no debida al factor de estudio si no al azar.

V R= $\dfrac{\sum x^2 - \left(n_1 \cdot m_1^2 + n_2 m_2^2 + n_3 \cdot m_3^2\right)}{N-k}$

Procedemos a la aplicación de la fórmula y con el objetivo de facilitar su observación calculamos algunos valores que indicamos abajo den la tabla:

Raza Blanca. Muestra 1	Raza negra. Muestra 2	Raza amarilla. Muestra 3
170	172	162
182	184	156
168	169	166
175	167	171
185	171	158
166	183	161
178	164	
180		
$n_1 = 8$	$n_2 = 7$	$n_3 = 6$
$m_1 = 175,5$	$m_2 = 172,86$	$m_3 = 162,33$
$\sum X^2 = 246.738$ cm.	$\sum X^2 = 209.516$ cm	$\sum X^2 = 158.262$ cm

Siendo n = el número de personas de cada muestra

Siendo m = la media de cada muestra

Siendo $\sum X^2$ = Sumatorio del cuadrado de cada una de las puntuaciones de la muestra (170 X170 + 182x 182+…….).

Entre los pasos a seguir debemos calcular:

1.- Debemos calcular $\sum X^2$= Sumatorio de los cuadrados de las tres muestras =

246.738+209.516+158.262= 614.516 cm.

2.- Calculamos k = el número de muestras existentes. En nuestro caso son 3.

3.- Calculamos N= Suma de n1 + n2+ n3. Es decir, la suma total de las personas de las tres muestras, que han participado en el estudio. En nuestro caso son 8+7+6= 21 individuos.

4.- Calculamos M, o media de las medias de todas las muestras. En nuestro caso las tres muestras.

M = $n_1m_1 + n_2m_2 + n_3m_3$ / N. En nuestro caso = 175,8x8+172.86x7+162,33x6/ 21 = 170.84.

5.- Debemos continuar calculando la varianza factorial

V F = $\dfrac{(n_1 \cdot m_1^2 + n_2 \cdot m_2^2 + n_3 \cdot m_3^2) - N \cdot M^2}{k-1}$

Siendo $(n_1 \cdot m_1^2 + n_2 \cdot m_2^2 + n_3 \cdot m_3^2)$. **Es decir por un lado la suma de cada uno de los productos de el número de personas que participaron en cada muestra y por otro la puntuación obtenida, elevada ésta al cuadrado. En nuestro caso = (8x 175,8+7x 172.86+ 6x 162,332) = 613.672,22 cm.** ^{22}
Por otro a este valor se le deduce NM. En nuestro caso 21 x 170,862 = 21x 129.123 = 613.055 ^{2}

6.- Finalmente se divide el valor obtenido en el numerador anterior por K-1. Siendo k el número de muestras que participan en nuestro estudio En nuestro caso 3-1 = 2

Siendo por tanto V F= 613.672,22- 613.055 / 2 = 308.15

7.- Debemos calcular

Debemos calcular ahora la varianza residual.

$$V R = \frac{\sum x^2 - \left(n_1 \cdot m_1^2 + n_2 m_2^2 + n_3 \cdot m_3^2\right)}{N-k}$$

Siendo N- K = el número de personas totales que participan en el estudio – el número de muestras. En nuestro caso = 21-3 = 18. En nuestro caso VR = 614.516-613.672,22/18 = 46,88

8.- Calculamos ahora la F = V F/ V R. En nuestro caso = 308.15/ 46,82 = 6,57

9.- Debemos calcular ahora los grados de libertad, tanto para la varianza factorial como para la residual.

- **Los grados de libertad de la varianza factorial se calculan, k-1. Es decir, el número de muestras menos 1. En nuestro caso 3-1 = 2**
- **Los grados de libertad de la varianza residual se calculan, N-k. Es decir, el número de sujetos en estudio- el numero de muestras. En nuestro caso 21-3 = 18**

10.- Decidimos el nivel de confianza que queremos asumir en el estudio. En nuestro caso 99% o error = 0,01

11.- Procedemos a observar en la tabla de la F para F (2.18): p<0.01 = 6

12.- Procedemos a la interpretación de los datos. Como F de nuestro estudio es de 6,57 > a F de la tabla = 6, rechazamos la hipótesis nula de no relación o independencia. Por tanto concluimos que con una probabilidad de error del 1% (p< 0,01) existe relación entre la raza de las personas y su estatura.

Bibliografía

- **Carrasco JL. El método estadístico en la investigación médica. Editorial Ciencia 3. 6ª Edición. 1995**
- **Rodríguez Miñon P. Estadística Aplicada a la Biología. Editorial UNED. 3º Edición. 1984.**
- **Polit Denise y Hungler Bernadette. Investigación científica en ciencias de la salud. Editorial McGraw-Hill Interamericana. 6ª edición. 2000.**

Análisis de datos en los estudios epidemiológicos V.
Prueba de Chi cuadrado y análisis de la varianza

Julia García Salinero