

Tablas de contingencia y contrastes χ^2

Homogeneidad (bondad del ajuste)

Grados de Biología sanitaria

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

7 de diciembre de 2017

Contraste de homogeneidad: el problema

Analizar si los niveles de un factor siguen cierta distribución.

Es decir, dadas

- Una muestra de un factor (v. cualit).
- Una distribución teórica de probabilidades para los niveles

¿son coherentes?

Ejemplo: Factor: piel del guisante:

- 1 Niveles: rugosa/lisa
- 2 Distribución teórica: 75 % - 25 %

Ejemplo: Factor: resultado de lanzar un dado

- 1 Niveles: 1, 2,..., 6
- 2 Distribución teórica: frecuencias relativas = $1/6$

Ejemplo: Aquí están los resultados de lanzar 5000 veces un dado

Resultado	1	2	3	4	5	6
Frecuencia	$o_1 = 811$	$o_2 = 805$	$o_3 = 869$	$o_4 = 927$	$o_5 = 772$	$o_6 = 816$

Y aquí tenemos los valores que deberíamos haber observado si el dado estuviera equilibrado $5000/6 \approx 833.3$

Resultado	1	2	3	4	5	6
Frecuencia	$e_1 = \frac{5000}{6}$	$e_2 = \frac{5000}{6}$	$e_3 = \frac{5000}{6}$	$e_4 = \frac{5000}{6}$	$e_5 = \frac{5000}{6}$	$e_6 = \frac{5000}{6}$

$$H_0 : \{\text{dado no cargado}\} = \left\{ \text{la probabilidad de cada uno de los valores es } \frac{1}{6} \right\}$$

$$H_1 : \{\text{dado cargado}\} = \left\{ \text{la probabilidad de al menos un valor no es } \frac{1}{6} \right\} \text{ (al menos 2)}$$

¿Son **significativas** las diferencias? ¿Se deben al azar?

Contraste de hipótesis χ^2 de homogeneidad (bondad del ajuste)

Sea X una variable aleatoria con función de densidad

Valor	x_1	x_2	$\dots$	x_k
Probabilidad	p_1	p_2	$\dots$	p_k

Dada una muestra de tamaño n con **valores observados**:

Valor	x_1	x_2	$\dots$	x_k	Total
Frecuencia	o_1	o_2	$\dots$	o_k	n

y dados los **valores esperados**

$$e_1 = n \cdot p_1, \quad e_2 = n \cdot p_2, \quad \dots, \quad e_k = n \cdot p_k$$

el estadístico

$$\Xi = \sum_{j=1}^k \frac{(o_j - e_j)^2}{e_j} \text{ sigue una distribución } \chi_{n-1}^2$$

siempre que $n > 30$, $e_j \geq 5 \forall j$ y que sea cierta

$$H_0 : \{\text{Los niveles de distribuyen según } p_1, \dots, p_k\}$$

El p-valor del contraste es $P(\chi_{n-1}^2 > \Xi)$

Ejemplo: con los valores observados para el dado resulta

$$\Xi = \sum_{i=1}^6 \frac{(o_i - e_i)^2}{e_i} = \frac{(o_1 - e_1)^2}{e_1} + \dots + \frac{(o_6 - e_6)^2}{e_6} =$$

$$\frac{(811 - 833)^2}{833} + \frac{(805 - 833)^2}{833} + \frac{(869 - 833)^2}{833} + \frac{(927 - 833)^2}{833} + \frac{(772 - 833)^2}{833} + \frac{(816 - 833)^2}{833} \approx 18.49$$

El p-valor del contraste es

$$P\left(\chi_5^2 > 18.49\right) = 0.0024$$

lo que nos lleva a rechazar H_0 en favor de H_1 .

Homogeneidad e independencia: similitudes

- Comparamos valores observados con valores esperados (si H_0 cierta)
- El estadístico de contraste es la suma de

$$\frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

extendido a todas las celdas de la tabla y sigue una χ^2

Homogeneidad e independencia: diferencias

- Independencia: hay 2 variables cualitativas.
- Homogeneidad: los valores observados y esperados (teóricos) se refieren a una única variable.

Ejemplo En 1865, Gregor Mendel postuló que en los descendientes que forman una generación, la proporción entre semillas lisas y rugosas debía ser de 3 a 1 que, en términos de probabilidades, significa

$$H_0 : \{p_{\text{lisa}} = 0.75, \quad p_{\text{rugosa}} = 0.25\}$$

Mendel obtuvo 7324 semillas clasificadas como

Forma de la semilla	lisa	rugosa	total
Frecuencia	$o_1 = 5474$	$o_2 = 1850$	7324

Los valores esperados eran

Forma de la semilla	lisa	rugosa	total
Frecuencia	$e_1 = 7324 \cdot \frac{3}{4} = 5493$	$e_2 = 7324 \cdot \frac{1}{4} = 1831$	7324

El valor del estadístico de contraste es

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(5474 - 5493)^2}{5493} + \frac{(1850 - 1831)^2}{1831} \approx 0.2629$$

y su p-valor ≈ 0.61 . No hay motivo para dudar de H_0