

Tablas de contingencia y contrastes χ^2

Independencia

Grado 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 independencia: el problema

Analizaremos la (posible) relación entre dos variables cualitativas

$$F \sim F$$

Terminología habitual

- Una variable cualitativa se suele llamar **factor**
- Las modalidades de cada factor se llaman **niveles** del factor

Ejemplos:

- 1 Factor: Nivel de actividad,
niveles: muy bajo, bajo, normal, alto, muy alto
- 2 Factor: Creencia religiosa
niveles: creyente, no creyente
niveles (alternativo): cristiano, musulmán, budista, hinduista

Contraste de independencia: el problema

- Cada individuos puede clasificarse de acuerdo a los dos factores.
- La **tabla de contingencia** resume la (doble) clasificación.

Ejemplo: En el mes de enero de 2013 el *Barómetro* del CIS recoge las respuestas de $n = 2452$ personas sobre sus creencias religiosas

	Hombres	Mujeres
Creyentes	849	1015
No creyentes	356	232

Resultados del CIS de Enero de 2013 discriminados por género [aquí*](#)

¿Hay más creyentes entre los hombre, o entres las mujeres?

En otras palabras

¿Son independientes los factores “género” y “ser creyente”?

Hemos eliminado 19 mujeres y 12 hombres que decidieron no contestar

Tablas de contingencia 2×2

Hipótesis

H_0 : {La religiosidad es independiente del género}

H_1 : {La religiosidad dependen del género}

Ojo: si p_m y p_h son las proporciones de mujeres y hombres creyentes, podríamos contrastar

$$H_0 : p_m = p_h \qquad H_1 : p_m \neq p_h$$

como una diferencia de proporciones (para tablas 2×2 es equivalente)
pero no funciona si uno de factores tiene más de dos niveles.

Estrategia: comparar

- Lo que hemos **observado** en realidad
- Lo que **esperamos** observar si H_0 fuera cierta

las **diferencias** ¿son **significativas**? necesitaremos un estadístico y su distribución de probabilidad

Tablas de contingencia 2×2

Valores observados es la tabla obtenida del CIS y **valores marginales**

	Hombres	Mujeres	Total
Creyentes	$o_{11} = 849$	$o_{12} = 1015$	1864
No creyentes	$o_{21} = 356$	$o_{22} = 232$	588
Total	1205	1247	2452

Usamos

$$o_{ij}$$

para el número de observaciones

- del nivel i del primer factor y
- el nivel j del segundo factor

Tablas de contingencia 2×2

Valores esperados si H_0 cierta: idem proporción de creyentes por géneros

- Estimar la proporción creyentes sin considerar el género

$$\frac{\text{total creyentes}}{\text{total individuos}} = \frac{1864}{2452} = \hat{p} \approx 0.7602$$

- Proporción NO creyentes $\hat{q} = 1 - \hat{p}$

- Valores esperados:

$$e_{11} = \text{total hombres} \cdot \hat{p} = 1205 \cdot \hat{p} \approx 916.04$$

$$e_{12} = \text{total mujeres} \cdot \hat{p} = 1247 \cdot \hat{p} \approx 948.96$$

se obtiene la **tabla de valores esperados**

	Hombres	Mujeres	Total
Creயentes	$e_{11} = 916.04$	$e_{12} = 948.96$	1864
No creயentes	$e_{21} = 288.96$	$e_{22} = 299.04$	588
Total	1205	1247	2452

Tablas de contingencia 2×2

Test de independencia

Estadístico χ^2 para una tabla de contingencia 2×2

Dada una tabla de contingencia 2×2 , con valores esperados e_{ij} y valores observados o_{ij} (para $i, j = 1, 2$), definimos el estadístico:

$$\Xi = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}}$$

Entonces, si $n > 30$ y $e_{ij} \geq 5$ para $i, j = 1, 2$, se tiene que

$$\Xi \sim \chi_1^2$$

El p-valor del contraste es $P(\chi_1^2 > \Xi)$.

Simetría: da lo mismo considerar $F_1 \sim F_2$ que $F_2 \sim F_1$

Ejemplo: en el caso que nos ocupa

$$\begin{aligned}\Xi &= \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}} = \\ &= \frac{(849 - 916.04)^2}{916.04} + \frac{(1015 - 948.96)^2}{948.96} + \frac{(356 - 289.96)^2}{289.96} + \frac{(232 - 299.04)^2}{299.04} \approx 40.23\end{aligned}$$

Como $P(\chi_1^2 > 40.23) = 2.26 \cdot 10^{-10}$ **rechazar H_0**

Tablas de contingencia $n_1 \times n_2$

Caso general Contrastar la (posible) relación $F_1 \sim F_2$ donde

- el factor F_1 tiene n_1 niveles
- el factor F_2 tiene n_2 niveles

Considerar todas las combinaciones posibles de los niveles de F_1 y F_2 da una tabla de contingencia $n_1 \times n_2$, con n_1 filas y n_2 columnas:

		Variable F_2		
Variable F_1		b_1	$\cdots$	b_{n_2}
	a_1	o_{11}	$\cdots$	o_{1n_2}
	$\vdots$		$\ddots$	
	a_{n_1}	$o_{n_1 1}$	$\cdots$	$o_{n_1 n_2}$

Sección 12.1.2 del libro

Tablas de contingencia $n_1 \times n_2$

Para obtener los valores esperados se calculan los valores marginales

Variable F_1	Variable F_2			Total
	b_1	$\dots$	b_{n_2}	
a_1	o_{11}	$\dots$	o_{1n_2}	o_{1+}
$\vdots$		$\ddots$		$\vdots$
a_{n_1}	$o_{n_1 1}$	$\dots$	$o_{n_1 n_2}$	$o_{n_1 +}$
Total	$o_{+ 1}$	$\dots$	$o_{+ n_2}$	$o_{++}=n$

- Calcular las proporciones observadas con independencia del factor F_2

$$\hat{p}_i = \frac{\text{total individuos nivel } a_i}{\text{total individuos}} = \frac{o_{i+}}{n}$$

- Calcular la frecuencia esperada para cada par de niveles de F_1 y F_2

$$e_{ij} = \hat{p}_i \cdot \text{total individuos nivel } b_j = \hat{p}_i \cdot o_{+j}$$

Tablas de contingencia $n_1 \times n_2$

Test de independencia Estadístico χ^2 para una tabla de contingencia $n_1 \times n_2$

Dada una tabla de contingencia $n_1 \times n_2$, con valores observados o_{ij} , y valores esperados e_{ij} , definimos el estadístico:

$$\Xi = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{\text{tabla}} \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

Es decir, sumamos un término para cada casilla de la tabla. Entonces, *mientras sea $n > 30$ y ninguno de los valores e_{ij} sea menor de 5*, el estadístico Ξ sigue una distribución χ_k^2 , con

$$k = (n_1 - 1)(n_2 - 1)$$

grados de libertad. El p-valor del contraste

$$H_0 : \{\text{El factor } F_1 \text{ es independiente del factor } F_2\}$$

$$\text{es } P(\chi_{(n_1-1)(n_2-1)}^2 > \Xi)$$