Métodos de Inferencia Estadística con Proporciones y Frecuencias

Tablas de contingencia

Una tabla de contingencia o tabla RxC es aquella con un número de filas (i) y columnas (j) que tiene como entradas de la tabla los valores A_1, A_2 … , B_1, B_2 … de dos variables cualitativas A y B. En las celdas se recogen las frecuencias absolutas, es decir el número de individuos de una muestra aleatoria *n* que presentan valores A_i y B_j.

Por ejemplo, supóngase que estudiamos la presencia de dos atributos A y B en los individuos de una muestra. Los atributos toman valores A_1, A_2, A_3, A_4, A_5 y B_1, B_2 respectivamente, obteniéndose experimentalmente el número de individuos que presentan valores A_i y B_j, es decir las *frecuencias observadas* o_{ij}:

	A_1	A_2	A_3	A_4	A_5
B_1	8	22	16	12	19
B_2	24	6	2	12	7

El tamaño muestral *n* será la suma de las frecuencias de todas las celdas, esto es $\sum_i \sum_j o_{ij}$. Si las variables A y B fueran independientes entonces el valor de las *frecuencias esperadas* o valores teóricos en las celdas sería igual a e_{ij}. Esta frecuencia esperada se obtiene dividendo el producto de las *frecuencias marginales* (suma de fila o suma de columna) entre el tamaño muestral *n*:

$$e_{ij} = \frac{o_i^* . o_j^*}{n}$$

Finalmente, utilizando el siguiente estadístico con distribución χ^2 y un número de grados de libertad igual a (filas-1).(columnas-1), se comparan las frecuencias observadas con las esperadas:

$$\chi^2 = \sum_i \sum_j \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

El caos descrito corresponde a lo que se denomina *test de independencia*, siendo el contraste:

H0: "A es independiente de B"

Ha: "A es dependiente de B"

Si la tabla de contingencia fuera una tabla Rx2 entonces el test se denomina *test de homogeneidad* contrastándose si las muestras B_1, B_2, … , B_k proceden o no de la misma población, es decir si hay o no diferencias significativas:

$$\text{H0}: p_1 = p_2 = \ldots = p_k$$

$$\text{Ha}: p_1 \neq p_2 \neq \ldots \neq p_k$$

siendo p_1, p_2, …, p_k la proporción de sujetos de las muestras 1, 2 y *k* que presentan el atributo A_1:

	A_1	A_2
B_1	8	22
B_2	24	6
B_3	…	…
…	…	…
B_k	…	…

Rafael Lahoz-Beltrá, Pilar López González-Nieto, Mariángeles Gómez Flechoso, María Eugenia Arribas, Mocoroa, Alfonso Muñoz Martín, María de la Luz García Lorenzo, Gloria Cabrera Gómez, Jose Antonio Alvarez Gómez, Andrea Caso Fraile, Jefferson Mark Orosco Dagan, Raul Merinero Palomar. Universidad Complutense de Madrid, 2017.