Análisis de Correspondencias (AC)

Hamdi Raïssi

IES PUCV

hamdi.raissi@pucv.cl

Objetivos de esta parte

Objetivos :

a- Repaso del test de independencia del χ^2.

b- Representar y comprender la estructura de dependencia entre variables cualitativas.

Los datos.

Tenemos dos variables cualitativas X y Y. Podemos representar las juntos en un cuadro de contingencia :

$X\backslash Y$	B_1	$\ldots$	B_j	$\ldots$	B_y	Total
A_1			n_{1j}			
$\vdots$			$\vdots$			
A_i	$\ldots$	$\ldots$	n_{ij}	$\ldots$	$\ldots$	$n_{i.}$
$\vdots$			$\vdots$			
A_x			$\vdots$			
Total			$n_{.j}$			n

probabilidades empiricas.

- n_{ij} : número de individuos con las modalidades A_i y B_j.
- n_{ij}/n : frecuencia de $A_i \cap B_j$ o estimación de $P(A_i \cap B_j)$.
- Si X y Y son independientes, $n_{i.}n_{.j}/n^2$ es también una estimación de $P(A_i \cap B_j) = P(A_i)P(B_j)$
- $n_{ij}/n_{i.}$ y $n_{ij}/n_{.j}$ nos dan las distribuciones condicionales (empiricas).
- Las últimas columna y fila *(divididas por n)* son las distribuciones marginales (empiricas).

Test de independencia del χ^2.

- La idea del test χ^2 : comparar n_{ij} con $n_{i.}n_{.j}/n$.
- Se puede mostrar que la estadística

$$\chi^2_{cal} = \sum_{i=1}^{x} \sum_{i=1}^{y} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n},$$

  sigue una ley $\chi^2_{(x-1)(y-1)}$.
- En R : `chisq.test(cuadro.contingencia)`.
- Considerar los bases de datos `UCBAdmissions`, `Titanic` y `HairEyeColor`.

Test de independencia del χ^2.

`library(datasets)`

- `chisq.test(HairEyeColor[,,2])` : test de independencia de colores de pelos y ojos de las mujeres (poner 1 para los hombres).
- `chisq.test(UCBAdmissions[,,6])` : test de independencia para el "Dept F" de la Universidad en termino de genero y rechazo/aceptación (poner 1,2,3,4,5 por los otros Dept).
- `chisq.test(Titanic[,,2,2])` : test de independencia entre genero y clase/tripulación de los adultos sobrevivientes del Titanic (poner 1 por los niños en la tercera coordenada y 1 por los fallecidos en la cuarta coordenada).
  Manipulando la base de datos se podría hacer un test de independencia fallecido/sobreviviente con clases y tripulación.

Test de independencia del χ^2 y análisis de correspondencias.

- Es una conclusión general, sin embargo queremos tener una imagen más precisa de las dependencias.
- Vamos a presentar el análisis de correspondencias (AC) por filas.
- La presentación es la misma por columnas, y se puede mostrar que las columnas y filas tienen un rol simetrico.

Análisis de correspondencias : intuición matematica.

- Si tenemos independencia :

$$\frac{P(A_i \cap B_j)}{P(A_i)} = P(B_j/A_i) = P(B_j).$$

- En practica (*"multiplicando" por* n *abajo y arriba la ecuación anterior a la izquierda*) :

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n},$$

o de manera *approximativa* en general...

Análisis de correspondencias : intuición matematica.

- Para cada modalidad de X, podemos definir los perfiles filas (es un vector en R^y) :

$$P_{A_i} = (n_{i1}/n_{i.}, n_{i2}/n_{i.}, \dots, n_{iy}/n_{i.})'.$$

- En caso de independencia tenemos de acuerdo a las ecuaciones de la diapositiva anterior :

$$P_{A_1} = P_{A_2} = \cdots = P_{A_x} = G_X,$$

o de manera approximativa en general con

$$G_X = (n_{.1}/n, n_{.2}/n, \dots, n_{.y}/n)'$$

Los perfiles :

$X \backslash Y$	B_1	$\dots$	B_j	$\dots$	B_y	Total
P_{A_1}			$n_{1j}/n_{1.}$			1
$\vdots$			$\vdots$			1
P_{A_i}	$\dots$	$\dots$	$n_{ij}/n_{i.}$	$\dots$	$\dots$	1
$\vdots$			$\vdots$			1
P_{A_x}			$n_{xj}/n_{x.}$			1
G_X			$n_{.j}/n$			1

Tenemos x puntos $(P_{A_1}, \dots, P_{A_x})$ que se ubican en $\mathbb{R}^y$.

Los perfiles :

- Todas las filas deben ser iguales en caso de independencia (o de manera approximativa).
- La manera de como los puntos se descartan entre ellos nos da información sobre la dependencia.
- A muchas veces tenemos muchas modalidades, los perfiles son en R^y, con $y >> 2$.
- Dificil de comprender como son ubicados los puntos en R^y....

Los perfiles :

- Vamos a proyectar la nube de puntos $P_{A_1}, P_{A_2}, \ldots, P_{A_x}$ de R^y sobre R^2 (en general).
- Minizando las distorciones del nube de puntos.
- Por eso ocupamos la distancia del chi-cuadrado entre dos perfiles :

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{y} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 .$$

Los perfiles :

- El termino del chi-cuadrado viene del hecho que la inercia (o varianza, o suma de las distancias al centro de gravedad) del nube de puntos es la estadística del test de independencia del χ^2.
- Cuando dos perfiles parecen similares (que sean columnas o filas) se justifica la fusión de las columnas.
- Utilizando $d^2_{\chi^2}(i,i')$, si se juntan dos perfiles columnas, la distancia entre los perfiles filas no cambian.

Los perfiles :

- Se optimiza la variabilidad de los nubes proyectados.
- Se obtiene vectores propios y valores propios tal que ACP
- Sin embargo los valores propios son entre 0 y 1.

Los perfiles :

- Hacer el AC de la base de datos `HairEyeColor`, hombre o mujeres (o los dos) como quieren...
- Se destaca la oposición `Blue/Blond` <——>`Black,Brown/Brown`
- Las otras modalidades no parecen bien representada en el AC.
- No hay un punto cerca el centro de gravidad (o sea modalidades "promedio" de la población). Hay muchos contrastes.

Como el AC describe la dependencia :

Comentarios del estudio :

- Más nos descartamos del centro de gravedad, más hay dependencia.
- Identificamos los perfiles que se alejan más del centro de gravedad.
- De la diferencias entre puntos obtenemos la interpretación de los componentes principales.

Como el AC describe la dependencia :

Estadística V (de Cramer) :

- Se puede mostrar que la inercia total χ^2_{cal} (la estadística del test de independencia) es tal que

$$\frac{\chi^2_{cal}}{n} \leq \min(x-1, y-1)$$

- Definimos :

$$V = \frac{\chi^2_{cal}}{n \min(x-1, y-1)}$$

- Estadística V (de Cramer) : si es cerca 0 estamos cerca la independencia. Si es cerca 1 estamos con relaciones entre variables fuerte.