Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

Análisis de Correspondencias

Test Chi Cuadrado

Definición

Una tabla de contingencia es una tabla que consiste de al menos dos filas y dos columnas para representar datos categóricos en términos de conteos o frecuencias.

Definición

Una tabla de contingencia es una tabla que consiste de al menos dos filas y dos columnas para representar datos categóricos en términos de conteos o frecuencias.

Veamos un ejemplo ...

TABLAS DE CONTINGENCIA

	Química	Economía	Literatura	Medicina	Paz	Física	**Total**
Canada	4	3	2	4	1	4	18
Francia	8	3	11	12	10	9	53
Italia	1	1	6	5	1	5	19
Alemania	24	1	8	18	5	24	80
Japón	6	0	2	3	1	11	23
Reino Unido	23	6	7	26	11	20	93
Estados Unidos	51	43	8	70	19	66	257
Rusia	4	3	5	2	3	10	27
Total	121	60	49	140	51	149	570

Cuadro 1: Premios Nobel repartidos entre los países del G8

TABLAS DE CONTINGENCIA

	Química	Economía	Literatura	Medicina	Paz	Física	**Total**
Canada	4	3	2	4	1	4	18
Francia	8	3	11	12	10	9	53
Italia	1	1	6	5	1	5	19
Alemania	24	1	8	18	5	24	80
Japón	6	0	2	3	1	11	23
Reino Unido	23	6	7	26	11	20	93
Estados Unidos	51	43	8	70	19	66	257
Rusia	4	3	5	2	3	10	27
Total	121	60	49	140	51	149	570

Cuadro 1: Premios Nobel repartidos entre los países del G8

¿Existe alguna relación entre países y categorías?

TABLAS DE CONTINGENCIA

	Química	Economía	Literatura	Medicina	Paz	Física	**Total**
Canada	4	3	2	4	1	4	18
Francia	8	3	11	12	10	9	53
Italia	1	1	6	5	1	5	19
Alemania	24	1	8	18	5	24	80
Japón	6	0	2	3	1	11	23
Reino Unido	23	6	7	26	11	20	93
Estados Unidos	51	43	8	70	19	66	257
Rusia	4	3	5	2	3	10	27
Total	121	60	49	140	51	149	570

Cuadro 1: Premios Nobel repartidos entre los países del G8

¿Existe alguna relación entre países y categorías?

¿Algunos países se especializan en ciertas áreas más que otros?

TABLAS DE CONTINGENCIA

Modelando el problema:

Tenemos una muestra de tamaño n y 2 variables cualitativas. Podemos resumir la información en una tabla,

$$
\begin{array}{cccccc|c}
n_{11} & n_{12} & \dots & n_{1j} & \dots & n_{1J} & \mathbf{n_{1.}} \\
n_{21} & n_{22} & \dots & n_{2j} & \dots & n_{2J} & \mathbf{n_{2.}} \\
\vdots & \vdots & & \vdots & & \vdots & \vdots \\
n_{i1} & n_{i2} & \dots & n_{ij} & \dots & n_{iJ} & \mathbf{n_{i.}} \\
\vdots & \vdots & & \vdots & \ddots & \vdots & \vdots \\
n_{I1} & n_{I2} & \dots & n_{IJ} & \dots & n_{IJ} & \mathbf{n_{I.}} \\
\hline
\mathbf{n_{.1}} & \mathbf{n_{.2}} & \dots & \mathbf{n_{.j}} & \dots & \mathbf{n_{.J}} & \mathbf{n_{..}} = \mathbf{n}
\end{array}
$$

Cuadro 2:

La cantidad de individuos en cada celda de la tabla tiene distribución multinomial, es decir se puede modelar por una tabla aleatoria $N \sim \mathcal{M}\left(n, [p_{ij}]_{ij}\right)$ donde p_{ij} es la probabilidad que una observación caiga en la fila i y la columna j.

TABLAS DE CONTINGENCIA

La cantidad de individuos en cada celda de la tabla tiene distribución multinomial, es decir se puede modelar por una tabla aleatoria $N \sim \mathcal{M}\left(n, [p_{ij}]_{ij}\right)$ donde p_{ij} es la probabilidad que una observación caiga en la fila i y la columna j.

Recordemos cómo se calculaban las marginales:

$$p_j = \sum_{i=1}^{I} p_{ij} \text{ probabilidad de caer en la columna } j.$$

$$p_i = \sum_{j=1}^{J} p_{ij} \text{ probabilidad de caer en la fila } i.$$

TABLAS DE CONTINGENCIA

p_{11}	p_{12}	$\dots$	p_{1j}	$\dots$	p_{1J}	$p_{1.}$
p_{21}	p_{22}	$\dots$	p_{2j}	$\dots$	p_{2J}	$p_{2.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$	p_{iJ}	$p_{i.}$
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
p_{I1}	p_{I2}	$\dots$	p_{IJ}	$\dots$	p_{IJ}	$p_{I.}$
p_{1}	p_{2}	$\dots$	p_{j}	$\dots$	p_{J}	1

Cuadro 3:

Ahora podemos escribir las hipótesis para un test,

$$H_0 : p_{ij} = p_i p_j \; \forall \; i, j$$
$$H_A : \exists \; i, j \text{ tal que } p_{ij} \neq p_i p_j$$

Ahora podemos escribir las hipótesis para un test,

$$
\begin{aligned}
H_0 &: p_{ij} = p_i p_j \; \forall \; i, j \\
H_A &: \exists \; i, j \text{ tal que } p_{ij} \neq p_i p_j
\end{aligned}
$$

Bajo H_0 el estimador de máxima verosimilitud de p_{ij} es

$$\hat{p}_{ij} = f_{i.} f_{.j},$$

donde,

$$f_{i.} = \frac{n_{i.}}{n} \;\text{ y }\; f_{.j} = \frac{n_{.j}}{n}$$

TABLAS DE CONTINGENCIA

Definimos las siguientes cantidades,

- **Observados:** $O_{ij} = n_{ij} = nf_{ij}$ la cantidad de individuos que caen en el lugar i, j de la tabla.
- **Esperados:** $E_{ij} = n\hat{p}_{ij} = \frac{n_{i.}n_{.j}}{n}$ la cantidad de individuos esperados bajo H_0 que caigan en el lugar i, j de la tabla.

Definimos las siguientes cantidades,

- **Observados:** $O_{ij} = n_{ij} = nf_{ij}$ la cantidad de individuos que caen en el lugar i, j de la tabla.
- **Esperados:** $E_{ij} = n\hat{p}_{ij} = \frac{n_{i.}n_{.j}}{n}$ la cantidad de individuos esperados bajo H_0 que caigan en el lugar i, j de la tabla.

Luego, el estadístico del test,

$$\chi = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Bajo H_0, $\chi \sim \chi^2_{(I-1)(J-1)}$ asintótica.

Observemos

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(nf_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2\frac{n}{n_{i.}n_{.j}} =$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(nf_{ij}-n\frac{n_{i.}n_{.j}}{n^2}\right)^2\frac{n}{n_{i.}n_{.j}} = \sum_{i=1}^{I}\sum_{j=1}^{J} n(f_{ij}-f_{i.}f_{.j})^2\frac{n^2}{n_{i.}n_{.j}}$$

$$\chi^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}(f_{ij}-f_{i.}f_{.j})^2\frac{1}{f_{i.}f_{.j}} = n\Phi^2$$

Φ^2 lleva el nombre de inercia.

Es posible tener una escritura matricial de la inercia.

Llamemos $N = [n_{ij}]$ y $P = \frac{N}{n} = [f_{ij}]$. Consideremos los vectores,

$$r = P \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} f_{1.} \\ \vdots \\ f_{I.} \end{pmatrix} \text{ y } c = P^t \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} f_{.1} \\ \vdots \\ f_{.J} \end{pmatrix}$$

Construimos a partir de r y c las matrices diagonales $D_r = diag(r)$ y $D_c = diag(c)$.

Observar

$$rc^t = \begin{pmatrix} f_{1.} \\ \vdots \\ f_{I.} \end{pmatrix} \begin{pmatrix} f_{.1} & \dots & f_{.J} \end{pmatrix} = \begin{pmatrix} f_{1.}f_{.1} & \dots & f_{1.}f_{.J} \\ \vdots & & \vdots \\ f_{I.}f_{.1} & \dots & f_{I.}f_{.J} \end{pmatrix}$$

$$P - rc^t = \begin{pmatrix} f_{11} - f_{1.}f_{.1} & \dots & f_{1J} - f_{1.}f_{.J} \\ \vdots & & \vdots \\ f_{I1} - f_{I.}f_{.1} & \dots & f_{IJ} - f_{I.}f_{.J} \end{pmatrix}$$

ESTRUCTURA MATRICIAL

Observar

$$rc^t = \begin{pmatrix} f_{1.} \\ \vdots \\ f_{I.} \end{pmatrix} \begin{pmatrix} f_{.1} & \dots & f_{.J} \end{pmatrix} = \begin{pmatrix} f_{1.}f_{.1} & \dots & f_{1.}f_{.J} \\ \vdots & & \vdots \\ f_{I.}f_{.1} & \dots & f_{I.}f_{.J} \end{pmatrix}$$

$$P - rc^t = \begin{pmatrix} f_{11} - f_{1.}f_{.1} & \dots & f_{1J} - f_{1.}f_{.J} \\ \vdots & & \vdots \\ f_{I1} - f_{I.}f_{.1} & \dots & f_{IJ} - f_{I.}f_{.J} \end{pmatrix}$$

$P - rc^t$ contiene en cada fila la diferencia entre la frecuencia observada y la frecuencia estimada bajo independencia.

Más aún, si hacen las cuentas, podemos expresar la inercia como

$$\Phi^2 = traza\left(D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t\right)$$

Más aún, si hacen las cuentas, podemos expresar la inercia como

$$\Phi^2 = traza \left(D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t)\right)$$

Objetivo: encontrar una representación en dimensión baja de las filas de $(P - rc^t)$ que resguarde la inercia lo más posible.

ESTRUCTURA MATRICIAL

Más aún, si hacen las cuentas, podemos expresar la inercia como

$$\Phi^2 = traza\left(D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t)\right)$$

Objetivo: encontrar una representación en dimensión baja de las filas de $(P - rc^t)$ que resguarde la inercia lo más posible.

Observación: la inercia juega el mismo papel que la variabilidad en el caso de componentes principales.

Para eso vamos a utilizar una técnica que se llama GSVD: Generalized Singular Value Descomposition.

GSVD

Para eso vamos a utilizar una técnica que se llama GSVD: Generalized Singular Value Descomposition.

¿De qué se trata?

GSVD

Para eso vamos a utilizar una técnica que se llama GSVD: **G**eneralized **S**ingular **Value D**escomposition.

¿De qué se trata?

Sean las matrices $\boldsymbol{\Omega} \in \mathbb{R}^{I \times I}$ y $\boldsymbol{\Psi} \in \mathbb{R}^{J \times J}$ simétricas y definidas positivas. Dada la matriz $A \in R^{I \times J}$ podemos factorizarla como

$$A = UD_\alpha V = \sum_{k=1}^{K} \alpha_k u_k v_k^t,$$

donde K es el rango de A, $U^t \boldsymbol{\Omega} U = I_{K \times K}$, $V^t \boldsymbol{\Psi} V = I_{K \times K}$ y $D_\alpha = diag(\alpha_1, \ldots, \alpha_K)$ con $\alpha_1 \geq \cdots \geq \alpha_K > 0$.

Observemos dos cosas importantes:

Observemos dos cosas importantes:

(1) Las columnas de U y V son ortogonales para el producto interno dado por Ω y Ψ.

Observemos dos cosas importantes:

(1) Las columnas de U y V son ortogonales para el producto interno dado por Ω y Ψ.

(2) La descomposición GSVD $A = \sum_{k=1}^{K} \alpha_k u_k v_k^t$ nos permite escribir las filas de la matriz A como combinación lineal de las columnas de V. Es decir, si llamamos $A_1, \ldots A_I$ a las filas de la matriz A,

$$
\begin{aligned}
A_1 &= \alpha_1 u_1^1 v_1 + \alpha_2 u_2^1 v_2 + \cdots + \alpha_K u_K^1 v_K \\
&\ldots \\
A_I &= \alpha_1 u_1^I v_1 + \alpha_2 u_2^I v_2 + \cdots + \alpha_K u_K^I v_K
\end{aligned}
$$

GSVD EN NUESTRO PROBLEMA

La técnica GSVD resuelve nuestro problemas, es cuestión de saber quién es quién.

- $A = P - rc^t$
- $\Omega = D_r^{-1}$
- $\Psi = D_c^{-1}$

GSVD EN NUESTRO PROBLEMA

La técnica GSVD resuelve nuestro problemas, es cuestión de saber quién es quién.

- $A = P - rc^t$
- $\Omega = D_r^{-1}$
- $\Psi = D_c^{-1}$

Aplico GSVD y consigo U, V tales que

$$P - rc^t = \sum_{k=1}^{K} \alpha_k u_k v_k^t.$$

Es decir, escribo las filas de $P - rc^t$ como combinación lineal de las columnas de V que son ortogonales con respecto al producto interno dado por D_c^{-1}.

GSVD EN NUESTRO PROBLEMA

Finalmente para tener una representación que se pueda dibujar truncamos la suma en $K^{\star} = 2$ o 3.

Finalmente para tener una representación que se pueda dibujar truncamos la suma en $K^\star = 2$ o 3.

¿Cómo controlamos lo que perdemos al truncar?

Es posible probar que

$$\Phi^2 = traza\left(D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t)\right) = \sum_{k=1}^{K} \alpha_k^2.$$

GSVD EN NUESTRO PROBLEMA

Finalmente para tener una representación que se pueda dibujar truncamos la suma en $K^\star = 2$ o 3.

¿Cómo controlamos lo que perdemos al truncar?

Es posible probar que

$$\Phi^2 = traza\left(D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t)\right) = \sum_{k=1}^{K} \alpha_k^2.$$

Procedemos parecido a las componentes principales comparamos la cantidad de inercia que juntamos truncando en $K^\star$.

GSVD EN NUESTRO PROBLEMA

El mismo análisis se puede hacer para las columnas.

Simplemente debemos transponer la matriz P y resulta que la inercia coincide.

COFFEE BREAK!