Tests du khi-deux
Homogénéité et indépendance

Mohamed LEMDANI

MISO
Université de Lille

30 Septembre 2021

Variable catégorielle

X variable catégorielle $\Longrightarrow$ prend un nombre fini de valeurs (modalités/catégories) :

$X = 1, 2, \ldots, c$

Exemples : groupe sanguin, nombre d'enfants, IMC (regroupée en classes), ...

Loi (distribution) de X :

Valeurs	1	2	...	c	
Probabilités	$\pi_1 = P(X = 1)$	π_2	...	π_c	$\sum_i \pi_i = 1$
Effectifs observés	O_1	O_2	...	O_c	$\sum_i O_i = n$

Types de tests portant sur des données catégorielles :

- Comparer la loi de X à une loi théorique (test d'ajustement).
- Comparer les lois de X entre les populations (test d'homogénéité).
- Tester l'indépendance de deux variables X et Y (test d'indépendance).

Test d'homogénéité : 1 variable catégorielle X et 2 (ou plusieurs) populations

$X = 1, 2, \ldots, c$ et l populations étudiées $\Longrightarrow$ l échantillons de $n_1, n_2, \ldots, n_l$ observations.

Tableau de contingence des effectifs observés O_{ij}

		X				
		1	2	$\ldots$	c	Totaux
	1	O_{11}	O_{12}	$\ldots$	O_{1c}	n_1
Échantillon	2	O_{21}	O_{22}	$\ldots$	O_{2c}	n_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	l	O_{l1}	O_{l2}	$\ldots$	O_{lc}	n_l
	Totaux	m_1	m_2	$\ldots$	m_c	n

$p_{11} = O_{11}/n_1$: proportion de $X = 1$ dans l'échantillon 1.

$p_{21}, \ldots, p_{l1}$: proportions de $X = 1$ dans les autres échantillons.

$p_1 = m_1/n$: proportion de $X = 1$ dans l'ensemble des échantillons.

$H_0 : \{\mathcal{L}_1 = \ldots = \mathcal{L}_l\}$ versus $H_1 : \{\text{il y a au moins deux lois distinctes}\}$.

H_0 vraie $\Longrightarrow p_{11} \approx p_{12} \approx \ldots \approx p_{l1} \approx p_1 \Longrightarrow O_{11}/n_1 \approx m_1/n \Longrightarrow O_{11} \approx \frac{n_1 \times m_1}{n} = T_{11}$.

T_{ij} : effectif *théorique* pour $X = i$ et $Y = j$ $\longrightarrow T_{ij} = \frac{n_i \times m_j}{n}$.

Test du khi-deux d'homogénéité (suite)

Variable de décision :

$$k = \sum_{ij} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{ij} \frac{O_{ij}^2}{T_{ij}} - n \sim \chi^2_{(l-1)(c-1)} \text{ sous } H_0.$$

Conditions : Tous les $T_{ij} \geqslant 5$.

Nombre de ddl : (Nombre de lignes -1) $\times$ (Nombre de colonnes-1).

Exemple 6 : On souhaite évaluer une nouvelle méthode pour combattre le stress. Pour cela, on la teste auprès d'un échantillon paritaire de 100 personnes. On note une amélioration auprès de 29 femmes (qui sont donc moins stressées) et une détérioration pour 11 autres (il n'y avait pas de différence notable pour les autres, entre avant et après). Pour les hommes, les chiffres respectifs étaient de 25 "améliorations" et de 15 "détériorations".

Peut-on dire que l'effet de la méthode diffère entre les hommes et les femmes au seuil de 10% ?

Exemple 6 (suite) :

Variable observée : X = 'Effet de la méthode' (A, D, =), deux échantillons

$n_F = n_H = 50$ et lois de X : π_A, π_D et $\pi_=$ pour les F et les H.

$H_0 : \{\mathcal{L}_F = \mathcal{L}_H\}$ *versus* $H_1 : \{\mathcal{L}_F \neq \mathcal{L}_H\}$

	A	D	=	Total
Femmes	29 (27)	11 (13)	10 (10)	50
Hommes	25 (27)	15 (13)	10 (10)	50
Total	54	26	20	100

$T_{11} = \dfrac{50 \times 54}{100} = 27,\ T_{12} = \dfrac{50 \times 26}{100} = 13, \ldots \Longrightarrow$ tous les $T_{ij} \geqslant 5$.

Variable de décision : $k = \displaystyle\sum_{cases} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} \sim \chi^2_{1\times 2} = \chi^2_2$ sous H_0.

Zone de rejet (10%) : $k_c \notin [4.605, +\infty[\Longrightarrow$ non rejet de H_0 au seuil de 10%.

Calculs : $k_c = \dfrac{(29-27)^2}{27} + \dfrac{(25-27)^2}{27} + \dfrac{(11-13)^2}{13} + \dfrac{(15-13)^2}{13} + 0 \approx 0.912.$

P-value : $0.5 < p < 0.7$.

Test d'indépendance : 2 variables catégorielles X et Y sur 1 population

$X = 1, 2, \ldots, l$ et $Y = 1, 2, \ldots, c$.

$H_0 : \{X, Y \text{ indépendantes}\}$ *versus* $H_1 : \{X, Y \text{ liées}\}$.

Tableau de contingence

		Y			
		1	2	...	c
	1	O_{11}	O_{12}	...	O_{1c}
X	2	O_{21}	O_{22}	...	O_{2c}
	⋮	⋮	⋮	⋱	⋮
	l	O_{l1}	O_{l2}	...	O_{lc}

Variable de décision :

$$k = \sum_{ij} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{ij} \frac{O_{ij}^2}{T_{ij}} - n \sim \chi^2_{(l-1)(c-1)} \text{ sous } H_0.$$

Conditions : Tous les $T_{ij} \geqslant 5$.

Conditions non remplies

Certains $T_{ij} < 5 \Longrightarrow$ regrouper (si cela a un sens) :

IMC	< 18	$[18, 25[$	$[25, 30[$	$[30, 40[$	$\geqslant 40$
T_i	7	23	18	9	3
Regroupement	7	23	18	12	

Cas d'un tableau 2×2 avec certains $T_{ij} < 5 \Longrightarrow$ khi-deux de Yates ou test exact de Fisher.