Statistique descriptive bidimensionnelle
Chapitre II : Deux variables qualitatives

Anas KNEFATI

Université Rennes 2

Plan

Plan

Présentation des données

Tableau de contingence

- X : Variable qualitative avec ℓ modalités : $x_1, x_2, ..., x_i, ..., x_\ell$
- Y : Variable qualitative avec c modalités : $y_1, y_2, ..., y_j, ..., y_c$

Présentation des données

Tableau de contingence

- X : Variable qualitative avec ℓ modalités : $x_1, x_2, ..., x_i, ..., x_\ell$
- Y : Variable qualitative avec c modalités : $y_1, y_2, ..., y_j, ..., y_c$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$						$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$						$\vdots$
x_ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell c}$	$n_{\ell .}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

- n_{ij} : Effectif conjoint de la ligne i et de la colonne j

Tableau de contingence

- X : Variable qualitative avec ℓ modalités : $x_1, x_2, ..., x_i, ..., x_\ell$
- Y : Variable qualitative avec c modalités : $y_1, y_2, ..., y_j, ..., y_c$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$						$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$						$\vdots$
x_ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell c}$	$n_{\ell .}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

- n_{ij} : Effectif conjoint de la ligne i et de la colonne j
- Effectif marginal de la ligne i : $n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$.

Tableau de contingence

- X : Variable qualitative avec ℓ modalités : $x_1, x_2, ..., x_i, ..., x_\ell$
- Y : Variable qualitative avec c modalités : $y_1, y_2, ..., y_j, ..., y_c$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$						$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$						$\vdots$
x_ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell c}$	$n_{\ell .}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

- n_{ij} : Effectif conjoint de la ligne i et de la colonne j
- Effectif marginal de la ligne i : $n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$.
- Effectif marginal de la colonne j : $n_{.j} = n_{1j} + n_{2j} + \cdots + n_{\ell j} = \sum_{i=1}^{\ell} n_{ij}$.

Tableau de contingence

- X : Variable qualitative avec ℓ modalités : $x_1, x_2, ..., x_i, ..., x_\ell$
- Y : Variable qualitative avec c modalités : $y_1, y_2, ..., y_j, ..., y_c$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$						$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$						$\vdots$
x_ℓ	$n_{\ell 1}$	$n_{\ell 2}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell c}$	$n_{\ell .}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

- n_{ij} : Effectif conjoint de la ligne i et de la colonne j
- Effectif marginal de la ligne i : $n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$.
- Effectif marginal de la colonne j : $n_{.j} = n_{1j} + n_{2j} + \cdots + n_{\ell j} = \sum_{i=1}^{\ell} n_{ij}$.
- Effectif total : $n = \sum_{i=1}^{\ell} \sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{\ell} n_{i.} = \sum_{j=1}^{c} n_{.j}$

Exemple : Femmes et hommes ont-ils les mêmes habitudes de lecture ?

Enquête faite sur les jeunes (16-24 ans) : INSEE 2012

- X : Sexe (F et M)
- Y : Taux de lecture de livres en 2012
  - $y = 0$: aucun
  - $y = 1$: moins de 6
  - $y = 2$: de 6 à moins de 12
  - $y = 3$: de 12 à moins de 24
  - $y = 4$: plus de 24

X \ Y	0	1	2	3	4	Total
F	27	36	22	9	5	99
H	54	23	12	6	4	99
Total	81	59	34	15	9	198

- $n_{1.} = 27 + 36 + 22 + 9 + 5 = 99$ et $n_{2.} = 54 + 23 + 12 + 6 + 4 = 99$
- $n_{.1} = 27 + 54 = 81$, $n_{2.} = 36 + 23 = 59$, etc
- $n = 99 + 99 = 198$

Plan

Fréquences conjointes et fréquences marginales

Définition

- Fréquence conjointe de la ligne i et et de la colonne j : $f_{ij} = \frac{n_{ij}}{n}$

Fréquences conjointes et fréquences marginales

Définition

- Fréquence conjointe de la ligne i et et de la colonne j : $f_{ij} = \frac{n_{ij}}{n}$
- Fréquence marginale de la ligne i : $f_{i.} = \frac{n_{i.}}{n}$
- Fréquence marginale de la colonne j : $f_{.j} = \frac{n_{.j}}{n}$

Fréquences conjointes et fréquences marginales

Définition

- Fréquence conjointe de la ligne i et et de la colonne j : $f_{ij} = \frac{n_{ij}}{n}$
- Fréquence marginale de la ligne i : $f_{i.} = \frac{n_{i.}}{n}$
- Fréquence marginale de la colonne j : $f_{.j} = \frac{n_{.j}}{n}$

Exemple : Habitudes de lecture

Table : Tableau des fréquences conjointes et fréquences marginales

X \ Y	0	1	2	3	4	$f_{i.}$
F	0.14	0.18	0.11	0.05	0.03	0.50
H	0.27	0.12	0.06	0.03	0.02	0.50
$f_{.j}$	0.41	0.30	0.17	0.08	0.05	1.00

$f_{13} = 0.11$, $f_{25} = 0.02$, $f_{2.} = 0.5$ et $f_{.1} = 0.41$

Plan

Définition

Profils lignes : Habitudes de lecture en France (2012)

Distribution conditionnelle de Y sachant que l'on est dans la modalité x_i de X : définie par les fréquences

$$f_{Y=y_j|X=x_i} = \frac{n_{ij}}{n_{i.}} \qquad \text{(notée} \quad f_{j|i})$$

X \ Y	0	1	2	3	4	Total
F	0.27	0.36	0.22	0.1	0.05	1
H	0.55	0.23	0.12	0.06	0.04	1

Définition

Profils lignes : Habitudes de lecture en France (2012)

Distribution conditionnelle de Y sachant que l'on est dans la modalité x_i de X : définie par les fréquences

$$f_{Y=y_j|X=x_i} = \frac{n_{ij}}{n_{i.}} \qquad \text{(notée} \quad f_{j|i})$$

X \ Y	0	1	2	3	4	Total
F	0.27	0.36	0.22	0.1	0.05	1
H	0.55	0.23	0.12	0.06	0.04	1

Profils colonnes : Habitudes de lecture en France (2012)

Distribution conditionnelle de X sachant que l'on est dans la modalité y_j de Y : définie par les fréquences

$$f_{X=x_i|Y=y_j} = \frac{n_{ij}}{n_{.j}} \qquad \text{(notée} \quad f_{i|j})$$

X \ Y	0	1	2	3	4
F	0.33	0.61	0.65	0.60	0.56
H	0.67	0.39	0.35	0.40	0.44
	1	1	1	1	1

Représentations graphiques

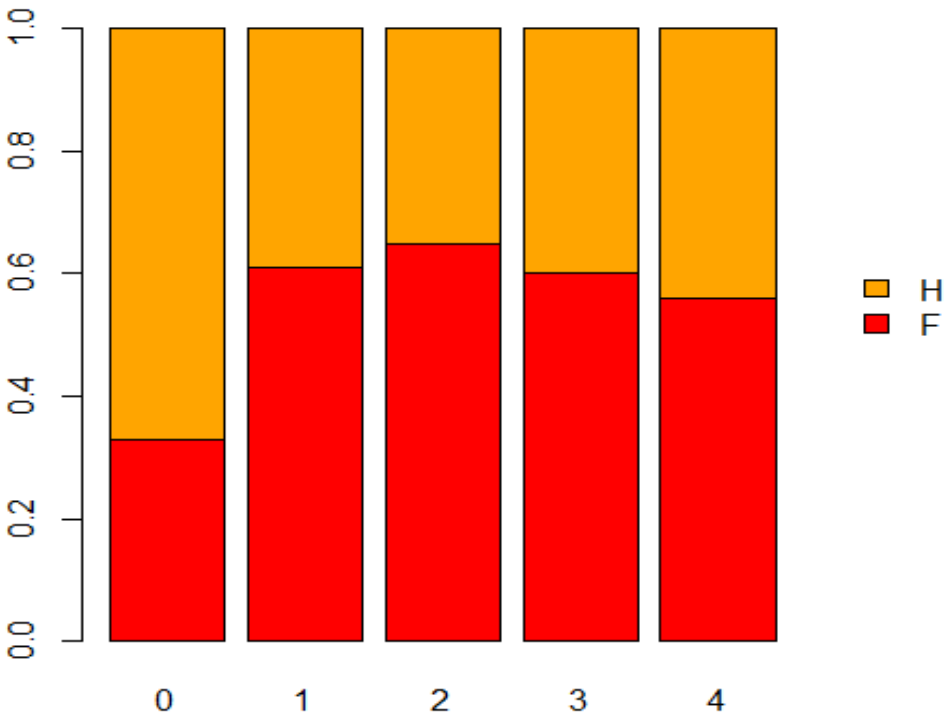

Figure : Profils-colonnes

Représentations graphiques

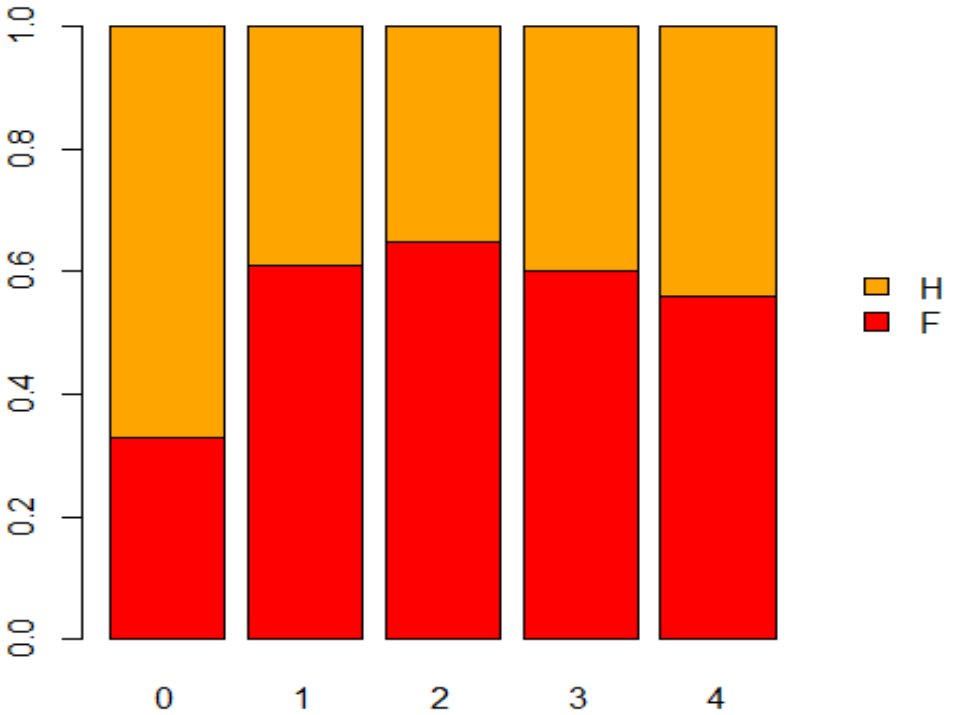

Figure : Profils-colonnes

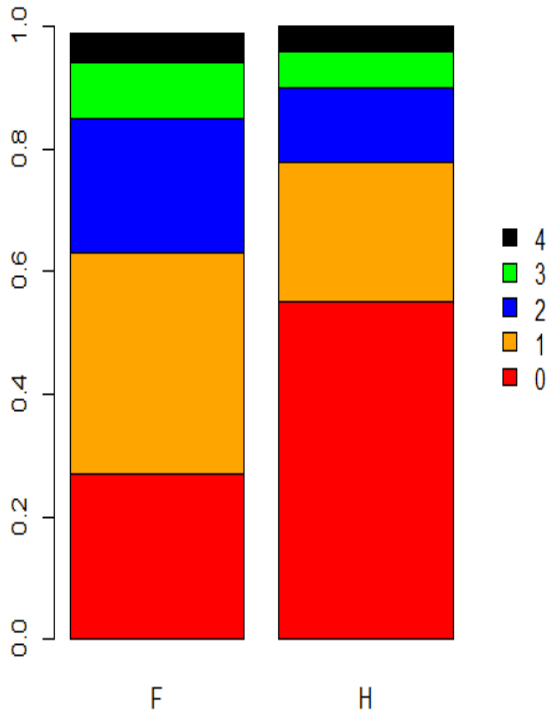

Figure : Profils-lignes

Plan

Khi deux : χ^2

Propriété

Les trois propriétés suivantes sont équivalentes :

- Tous les profils-lignes sont identiques ;
- Tous les profils-colonnes sont identiques ;
- Pour tout couple d'indices (i, j), on a

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} \quad \text{ou encore} \quad n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

Tableau d'indépendance théorique

- Ses éléments sont les effectifs théoriques : $n^*_{ij} = \frac{n_{i.} \times n_{.j}}{n}$.
- Ce tableau correspond à l'absence de lien entre les deux variables qualitatives X et Y
- Le khi deux : $\chi^2 = \sum_{i=1}^{\ell} \sum_{j=1}^{c} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$

Exemple

Tableau d'indépendance théorique : $n^*_{ij} = \frac{n_{i.} \times n_{.j}}{n}$.

	0	1	2	3	4
F	27	36	22	9	5
H	54	23	12	6	4

Table 2 : Effectifs observés

Exemple

Tableau d'indépendance théorique : $n_{ij}^* = \frac{n_{i.} \times n_{.j}}{n}$.

	0	1	2	3	4
F	27	36	22	9	5
H	54	23	12	6	4

Table 2 : Effectifs observés

	0	1	2	3	4
F	40.5	29.5	17	7.5	4.5
H	40.5	29.5	17	7.5	4.5

Table 1 : Effectifs théoriques

Exemple

Tableau d'indépendance théorique : $n_{ij}^* = \frac{n_{i.} \times n_{.j}}{n}$.

	0	1	2	3	4
F	27	36	22	9	5
H	54	23	12	6	4

Table 2 : Effectifs observés

	0	1	2	3	4
F	40.5	29.5	17	7.5	4.5
H	40.5	29.5	17	7.5	4.5

Table 1 : Effectifs théoriques

Tableau des valeurs partielles : $\frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$

	0	1	2	3	4
F	4.5	1.43	1.47	0.3	0.06
H	4.5	1.43	1.47	0.3	0.06

Alors $\chi^2 = 4.5 + 1.43 + 1.47 + ... + 0.06 = 15.52$

Exemple : Tableau des contributions au χ^2

Tableu des contributions : Valeur partielle/χ^2

	0	1	2	3	4
F	0.29	0.09	0.09	0.03	0
H	0.29	0.09	0.09	0.03	0

Exemple : Tableau des contributions au χ^2

Tableu des contributions : Valeur partielle/χ^2

	0	1	2	3	4
F	0.29	0.09	0.09	0.03	0
H	0.29	0.09	0.09	0.03	0

Tableau des contributions en pourcentage : 100 $\times$ Valeur partielle/χ^2

	0	1	2	3	4
F	28.99	9.21	9.47	1.93	0.4
H	28.99	9.21	9.47	1.93	0.4

Propriétés

- $\chi^2 \geq 0$
- $\chi^2 = n \left[\left(\sum_{i=1}^{\ell} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.} \times n_{.j}} \right) - 1 \right]$
- $\chi^2 = 0$ ssi n est dans un cas d'indépendance stricte
- χ^2 est d'autant plus grand que la liaison entre les deux variables considérées est plus forte
- $\chi^2 \leq n \times \min(\ell - 1, c - 1)$, cela signifie que χ^2 dépend de n, ℓ et c, ce qui est gênant pour l'interprétation concrète de ce coefficient.

Autre indicateurs liés χ^2

Coef de Pearson

$\Phi^2 = \frac{\chi^2}{n}$, c'est un coefficient indépendant de n

Coef de Tschuprow

$T = \sqrt{\frac{\Phi^2}{\sqrt{(\ell-1)(c-1)}}}$, c'est indépendant de n, ℓ et c

Coef de Cramer

$C = \sqrt{\frac{\Phi^2}{\min(\ell,c)-1}}$, c'est aussi indépendant de n, ℓ et c

Remarques sur T et C

- $0 \leq T \leq C \leq 1$
- T et C sont d'autant plus grands que la liaison entre les deux variables considérées est forte
- Dans la pratique, T et C sont rarement supérieurs à 0.5.

Exemple : Habitudes de lecture en France (2012)

- $\Phi^2 = \frac{\chi^2}{n} = \frac{15.52}{198} \approx 0.08.$
- $T = \sqrt{\frac{\Phi^2}{\sqrt{(\ell-1)(c-1)}}} = \sqrt{\frac{0.08}{\sqrt{(2-1)(5-1)}}} \approx 0.2$
- $C = \sqrt{\frac{\Phi^2}{\min(\ell,c)-1}} = \sqrt{\frac{0.08}{\min(2,5)-1}} \approx 0.28$

Comme C (ou T) est proche de zéro, alors la dépendance entre le taux de lecture et le sexe est faible.