Part II

Distributions statistiques bivariées

2 Distributions

2.1 Distributions conjointe, marginale, conditionnelle

Notons $m_1^X, \cdots, m_J^X$ les J modalités de X et $m_1^Y, \cdots, m_K^Y$ les K modalités de Y. Si l'une des deux variables (ou les deux) est quantitative continue, les m_j^X ou les m_k^Y sont des classes modales. Introduisons les quantités suivantes :

- n_{jk} est le nombre de fois où le couple (X, Y) prend la modalité (m_j^X, m_k^Y),
- $n_{\bullet k}$ est le nombre de fois où la variable Y prend la valeur m_k^Y,
- $n_{j\bullet}$ est le nombre de fois où la variable X prend la valeur m_j^X.

âge

nombre d'enfant.

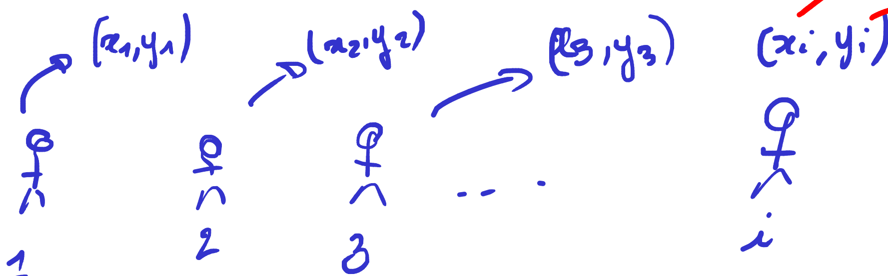


Tableau de contingence

modalités de Y

X \ Y	0	1	2	3	
[15, 20[	30	5	1	0	36
[20,30[	10	10	10	10	40
[30, 90[	5	20	10	10	45
	45	35	21	20	N = 121

effectifs conjoints

45 personnes ont entre 30 et 90 ans

loi marginale de X = "X sans Y".

modalités de X

35 personnes ont exactement 1 enfant

loi marginale de Y = "Y sans X".

Zoom sur 2ème ligne :

	0	1	2	3	
[20,30[	10	10	10	10	40

121

0	1	2	3
$\frac{10}{40}$	$\frac{10}{40}$	$\frac{10}{40}$	$\frac{10}{40}$

profil ligne

proportion de "0 enfant" parmi les [20, 30[
= fréquences conditionnelles des enfants sachant $X = [20, 30[$.

$\frac{10}{121}$	$\frac{10}{121}$	$\frac{10}{121}$	$\frac{10}{121}$

→ proportion des [20, 30["et" qui n'ont pas d'enfant dans la population totale.

On a

$$\sum_{j=1}^{J} n_{jk} = n_{\bullet k} \quad \text{et} \quad \sum_{k=1}^{K} n_{jk} = n_{j\bullet}$$

$$\sum_{k=1}^{K}\sum_{j=1}^{J} n_{jk} = \sum_{j=1}^{J} n_{j\bullet} = \sum_{k=1}^{K} n_{\bullet k} = n$$

n_{jk}

ligne j — colonne k

Les données peuvent être représentées dans un tableau à double entrée appelé $\boxed{\textbf{Tableau de contingence}}$.

ordre : ligne puis colonne

$X \backslash Y$	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
m_1^X	n_{11}	$\dots$	n_{1k}	$\dots$	n_{1K}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_j^X	n_{j1}	$\dots$	n_{jk}	$\dots$	n_{jK}	$n_{j\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_J^X	n_{J1}	$\dots$	n_{Jk}	$\dots$	n_{JK}	$n_{J\bullet}$
	$n_{\bullet 1}$	$\dots$	$n_{\bullet k}$	$\dots$	$n_{\bullet K}$	n

somme sur ligne J

$$n_{J\bullet} = \sum_{k=1}^{K} n_{Jk}$$

somme sur colonne 1.

$$n_{\bullet 1} = \sum_{j=1}^{J} n_{j1}$$

Le **tableau des fréquences** s'obtient en divisant les effectifs par le nombre d'unités statistiques n (effectif total). Comme précédemment on obtient

$$f_{jk} = \frac{n_{jk}}{n}, \quad f_{\bullet k} = \frac{n_{\bullet k}}{n} \quad f_{j\bullet} = \frac{n_{j\bullet}}{n}$$

$f_{1\bullet} = \sum_{k=1}^{K} f_{1k}$

	m_1^Y	$\ldots$	m_k^Y	$\ldots$	m_K^Y	total
m_1^X	f_{11}	$\ldots$	f_{1k}	$\ldots$	f_{1K}	$f_{1\bullet}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_j^X	f_{j1}	$\ldots$	f_{jk}	$\ldots$	f_{jK}	$f_{j\bullet}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_J^X	f_{J1}	$\ldots$	f_{Jk}	$\ldots$	f_{JK}	$f_{J\bullet}$
	$f_{\bullet 1}$	$\ldots$	$f_{\bullet k}$	$\ldots$	$f_{\bullet K}$	1

marges ou distributions marginales.

	m_1^Y	$\ldots$	m_k^Y	$\ldots$	m_K^Y	total
m_1^X	n_{11}	$\ldots$	n_{1k}	$\ldots$	n_{1K}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_j^X	n_{j1}	$\ldots$	n_{jk}	$\ldots$	n_{jK}	$n_{j\bullet}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_J^X	n_{J1}	$\ldots$	n_{Jk}	$\ldots$	n_{JK}	$n_{J\bullet}$
	$n_{\bullet 1}$	$\ldots$	$n_{\bullet k}$	$\ldots$	$n_{\bullet K}$	n

$\div n$

2.2 Distributions marginales

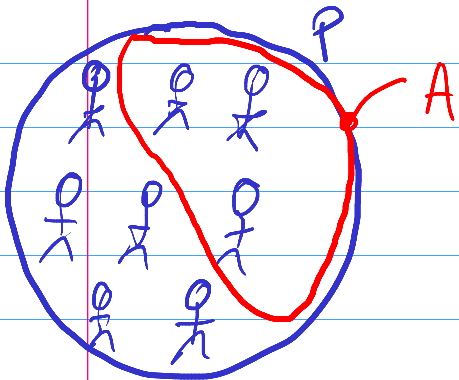
A partir du tableau de contingence, on peut retrouver la distribution de chacune des variables séparément :

Modalité de Y	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
Fréquence empirique	$f_{\bullet 1}$	$\dots$	$f_{\bullet k}$	$\dots$	$f_{\bullet K}$	1

Modalité de X	m_1^X	$\dots$	m_j^X	$\dots$	m_J^X	total
Fréquence empirique	$f_{1 \bullet}$	$\dots$	$f_{j \bullet}$	$\dots$	$f_{J \bullet}$	1

* Une fréquence : la fréquence de A dans P

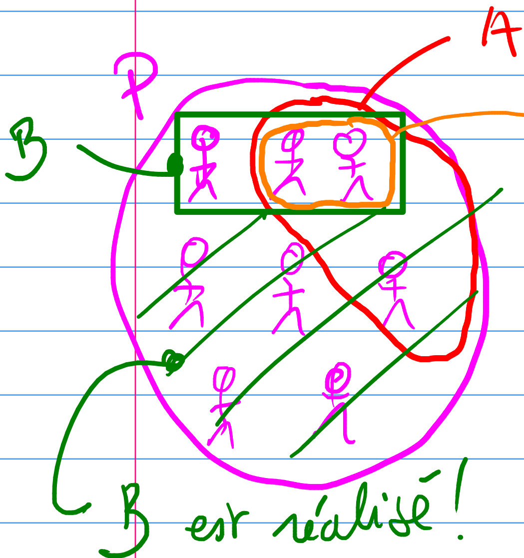
= la proportion d'individus A dans la populat° P.



$$f_{A|P} = \frac{n_A}{n_P} = \frac{3}{8}$$

Fréquence conditionnelle de A $\boxed{\text{sachant B}}$

B est réalisé !



$$f_{A|B} = \frac{n_{AB}}{n_B}$$

ce qu'il reste de A dans la population B.

$$f_{A|B} = \frac{2}{3}$$

$$f_{A|B} = \frac{n_{AB}}{n_B} = \frac{n_{AB}}{n_B} \times \frac{n}{n} = \frac{f_{AB}}{f_B}$$

f_{AB} $\qquad \frac{1}{f_B}$

Dans notre cadre :

effectif joint

$$f_{x=j|y=k} = \frac{n_{jk}}{n_{\bullet k}} = \frac{f_{jk}}{f_{\bullet k}}$$

$y=k$

" sachant que "

3.2 Exemple

À partir de 200 dossiers d'une agence immobilière, on recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42

92
108

100 100

Tableau des fréquences
% 200

	Cé	Co	
Acc.	0,17	0,29	0,46
Ref.	0,33	0,25	0,54
	0,5	0,5	

marges

3.2 Exemple (suite)

À partir de 200 dossiers d'une agence immobilière, on recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42

92

Distribution sachant dossier accepté

	Célibataire	En couple
Dossier accepté	$34/92 = 0{,}37$	$58/92 = 0{,}63$

⚠ fréquences conditionnelles.

S'appelle un profil ligne

3.2 Exemple

À partir de 200 dossiers d'une agence immobilière, on recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42
	100	

fréquences conditionnelles sachant Célibataire

	Célibataire
Accepté	$34/100 = 0{,}34$
Refusé	$66/100 = 0{,}66$

Profil colonne

2.3 Distributions conditionnelles

La ligne j du tableau de contingence représente la répartition sur les modalités (ou classes modales) $(m_1^Y, \cdots, m_K^Y)$ des individus pour lesquels le caractère X vaut m_j^X. Si on divise les lignes ou les colonnes par leur somme, on obtient les distributions empiriques constituées des fréquences conditionnelles. Pour $j = 1, \cdots, J$ et $k = 1, \cdots, K$ notons :

$$f_{k|j} = \frac{n_{jk}}{n_{j\bullet}} = \frac{f_{jk}}{f_{j\bullet}}.$$

La fréquence $f_{k|j}$ peut se lire fréquence de la modalité m_k^Y sachant que X prend la modalité m_j^X.

On peut alors construire le tableau des profils ligne :

$f_{1k}/f_{1\bullet}$

	m_1^Y	$\ldots$	m_k^Y	$\ldots$	m_K^Y	total
m_1^X	$f_{1\|1}$	$\ldots$	$f_{k\|1}$	$\ldots$	$f_{K\|1}$	1
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_j^X	$f_{1\|j}$	$\ldots$	$f_{k\|j}$	$\ldots$	$f_{K\|j}$	1
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_J^X	$f_{1\|J}$	$\ldots$	$f_{k\|J}$	$\ldots$	$f_{K\|J}$	1

$f_{kj}/f_{\bullet j}$

Les profils colonnes sont les fréquences en colonne i.e. :

$$f_{j|k} = \frac{n_{jk}}{n_{\bullet k}} = \frac{f_{jk}}{f_{\bullet k}}.$$

3 Quantification de la dépendance

3.1 Statistique du χ^2

En présence de deux variables, l'un des enjeux principaux est d'étudier (c'est à dire quantifier voire expliquer) la dépendance entre les deux caractères.

Indépendance pourrait se définir en disant que le fait que A se réalise n'influense pas la réalisation de B.

$$f_{B|A} = f_B$$

a $\quad f_{B|A} = f_B$

$\iff f_A f_{B|A} = f_B f_A$

$\iff \cancel{f_A} \dfrac{f_{B,A}}{\cancel{f_A}} = f_A f_B$

$\iff f_{BA} = f_A f_B$

la fréquence jointe est égale au produit des fréquences marginales.

En stat, cette notion d'indépendance est trop restrictive. Elle tomberait même pour un seul individu "mal placé" dans le tableau.

On regarde donc une distance au cadre d'indépendance ($f_{jk} = f_{j\cdot} f_{\cdot k}$).

$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\cdot} f_{\cdot k})^2}{f_{j\cdot} f_{\cdot k}}$$

ne pas oublier de multiplier par l'effectif total.

évite une trop forte contribution d'une modalité à effectif trop faible.

multiplié par n

somme

fréquence jointe

un carré pour le numérateur uniquement

En statistiques, on ne peut que “quantifier la distance à l’indépendance” par la statistique du χ^2,

$$
\begin{aligned}
D_{\chi^2} &= n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{\bullet j} f_{k \bullet})^2}{f_{\bullet j} f_{k \bullet}} \\
&= n \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{n_{jk}^2}{n_{j \bullet} n_{\bullet k}} - 1 \right)
\end{aligned}
$$

produit des marges.

Ou de façon équivalente :

$$
D_{\chi^2} = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{\bullet j} n_{k \bullet}}{n}\right)^2}{\frac{n_{\bullet j} n_{k \bullet}}{n}},
$$

où J et K sont le nombre de modalités de chacune des deux variables considérées.

Exemple :

$n = 200$

(i) On donne le tableau des fréquences.
Pour calculer les fréquences, on divise chaque effectif par l'effectif total (ici 200) :

	Célibataire	En couple	Total
Dossier accepté	0.17	0.29	0.46
Dossier refusé	0.33	0.21	0.54
Total	0.5	0.5	1

$$D_{\chi^2} = 200 \left(\frac{(0{,}17 - 0{,}5 \times 0{,}46)^2}{0{,}5 \times 0{,}46} + \frac{(0{,}29 - 0{,}46 \times 0{,}5)^2}{0{,}46 \times 0{,}5} + \frac{(0{,}33 - 0{,}5 \times 0{,}54)^2}{0{,}5 \times 0{,}54} + \frac{(0{,}21 - 0{,}54 \times 0{,}5)^2}{0{,}54 \times 0{,}5} \right)$$

$$
\begin{aligned}
D_{\chi^2} =& 200\left(\frac{(0.17-0.46\times 0.5)^2}{0.46\times 0.5}+\frac{(0.29-0.46\times 0.5)^2}{0.46\times 0.5}+\frac{(0.33-0.54\times 0.5)^2}{0.54\times 0.5}\right.\\
&\left.+\frac{(0.21-0.54\times 0.5)^2}{0.54\times 0.5}\right)\\
=& 200\Big(0.016+0.016+0.013+0.013\Big)\\
D_{\chi^2} =& 11.6
\end{aligned}
$$

3.3.1 Coefficients ϕ et C

Les coefficients ϕ et C découlent de la statistique du χ^2 par les formules

$$
C=\sqrt{\frac{D_{\chi^2}}{D_{\chi^2}+n}}, \qquad \phi=\sqrt{\frac{D_{\chi^2}}{n}}.
$$

En réalité ces deux coefficients sont une variante l'un de l'autre. L'avantage de C est qu'il est compris entre 0 et 1, alors que ce n'est pas le cas pour le ϕ. Plus ces indicateurs sont proche de zéro, plus il y a indépendance entre les deux variables X et Y étudiées.

3.3.2 V de Cramér

Comme pour le coefficient ϕ, plus le V de Cramér est proche de zéro, plus il y a indépendance entre les deux variables X et Y étudiées. Il vaut 1 en cas de complète dépendance.

Le coefficient V de Cramér nécessite l'utilisation de la statistique du χ^2 via la formule

$$V = \sqrt{\frac{D^2_\chi}{n \times min\{l-1; c-1\}}},$$

où n est l'effectif total de la population, c est le nombre de colonnes (nombre de modalités de Y) et l le nombre de lignes (modalits de X).

3.2 Exemple

À partir de 200 dossiers d'une agence immobilière, on recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42

$n = 200$

2 lignes
2 colonnes

$D_{\chi^2} = 11,6$

$$\min(l-1, c-1)$$
$$= \min(2-1, 2-1) = \underline{1}$$

$$V = \sqrt{\frac{D_{\chi^2}}{200 \times \min(l-1, c-1)}} = \sqrt{\frac{11,6}{200 \times \underline{1}}} = 0,24$$

3.3.3 Interprétation

L'interprétation des coefficients ϕ et V est empirique et dépend du domaine d'application (sciences économiques, sciences humaines, médecine...). On peut considérer le tableau suivant pour l'interprétation (tout en vérifiant les valeurs frontières d'usage dans chaque domaine).

Valeur du V de Cramér	Intensité de la relation entre les variables
inférieur à 0,10	relation nulle ou très faible
entre 0,10 et 0,20	relation faible
entre 0,20 et 0,30	relation moyenne
au dessus de 0,30	relation forte

la relation entre statut marital et acceptation du dossier est moyenne.