Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

Cours 3
Distributions conditionnelles

Partition de l'échantillon conditionnée par une variable

1 L'observation d'une population par une variable X conduit naturellement à un découpage de cette population, en regroupant ensemble tous les individus ayant la même modalité comme mesure pour X ; on obtient ainsi autant de sous-populations qu'il y a de modalités (k), chaque sous-population étant identifiée par la modalité commune à ses individus.

L'observation d'une variable X sur un *échantillon* conduit à un découpage similaire : en regroupant tous les individus qui ont une même modalité comme valeur, on obtient k sous-échantillons identifiés par cette modalité.

Dans le contexte d'une observation conjointe de deux variables X et Y, il y a deux découpages possibles, selon qu'on utilise l'une ou l'autre des deux variables. À partir de la situation « niveau scolaire et absentéisme » qu'on utilisera comme exemple dans ce cours, on peut découper l'échantillon avec la variable X « niveau scolaire », en identifiant deux sous-échantillons, celui des élèves de niveau A et celui des élèves de niveau B ; on peut également découper l'échantillon avec la variable Y « absentéisme » et obtenir 3 sous-échantillons, ceux des élèves rarement, moyennement et fréquemment absents.

Lorsqu'on étudie les liens entre deux variables à partir de leur observation conjointe sur un échantillon, il est assez naturel de comparer les distributions de l'une des variables sur les sous-échantillons créés par découpage à partir de l'autre : en reprenant l'exemple précédent, on voudra comparer les distributions de l'absentéisme sur les deux groupes de niveau scolaire pour vérifier l'hypothèse a priori selon laquelle il est globalement supérieur dans le groupe de plus faible niveau ; on pourrait tout aussi bien comparer les distributions du niveau scolaire sur les trois sortes d'absents pour vérifier si effectivement le niveau est globalement supérieur dans le groupe des élèves les plus assidus.

Nous allons donc consacrer ce chapitre à l'étude de ces distributions particulières appelées distributions conditionnelles : elles sont essentielles pour décrire la notion de liaison entre deux variables, objet du chapitre suivant.

2 déf On appelle sous-échantillon conditionné (induit) par la modalité m_i de X l'ensemble des individus de l'échantillon dont la mesure par X est m_i ; on le note $E_{x=m_i}$ ou plus simplement E_{m_i} ; avec k modalités, X conditionne donc k sous-échantillons, certains pouvant être vides.

Puisque les individus de E_{m_i} ont tous la modalité m_i, ce sont ceux qui participent aux effectifs de la ième ligne du tableau de contingence ; il y en a donc $n_{i.}$, le total de cette ligne : les tailles des k sous-échantillons induits sont donc les nombres de la distributions marginale de X en effectif.

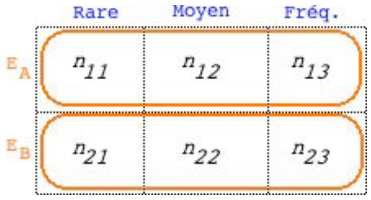
Exemple : dans la situation « niveau scolaire et absentéisme » il y a deux sous-échantillons induits par la variable X niveau scolaire : le sous-échantillon des élèves de niveau A, E_A, de taille 15, et celui des élèves de niveau B, E_B, de taille 12.

3 déf Les k sous-échantillons conditionnés par X forment une partition [1] de E, puisque chaque individu de l'échantillon appartient à un groupe (il a une valeur pour X) et un seul (il n'en a qu'une) ; cette partition $E_{m_1}, E_{m_2}, ..., E_{m_k}$ s'appelle partition de l'échantillon E conditionnée par X.

Exemple : dans la situation « niveau scolaire et absentéisme », la partition conditionnée par X

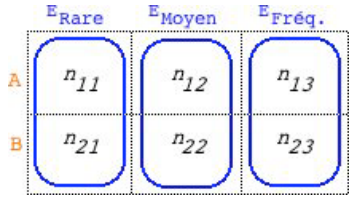
1. Une partition d'un ensemble E est un ensemble de parties de E, deux à deux disjointes (elles n'ont aucun élément commun) et dont la réunion est E tout entier ; cela est équivalent à dire que tout élément de E appartient à une partie et une seule.

se compose des deux sous-échantillons E_A et E_B (aucun élève ne peut être à la fois de niveau A et B, et chaque élève de l'échantillon est dans l'un des deux).



4 De manière symétrique, on appelle sous-échantillon conditionné par une modalité m'_j de Y, l'ensemble des $n_{.j}$ individus de l'échantillon dont la mesure par Y est m'_j ; on le note $E_{y=m'_j}$ ou plus simplement $E_{m'_j}$. Les p sous-échantillons conditionnés par Y, $E_{m'_1}, E_{m'_2}, ..., E_{m'_p}$, compose la partition de E conditionnée par Y.

Exemple : la variable absentéisme partitionne l'échantillon E en 3 sous-échantillons induites, E_{rare} E_{moyen} et $E_{frquent}$, de taille 15 6 et 6, composées des élèves rarement, moyennement ou fréquemment absents.



Distributions conditionnelles

5 déf Dans le contexte d'une observation conjointe XxY, on appelle distribution conditionnelle de X la distribution de X restreinte à un des sous-échantillon conditionnés par une modalité de Y : c'est la distribution de X quand on limite son observation aux individus ayant la même modalité pour Y ; il y en donc p ; par exemple la distribution de X sur le sous-échantillon $E_{m'_j}$ est la distribution de X conditionnée par m'_j, notée $X_{m'_j}$, ou encore la distribution conditionnelle de X sachant que Y vaut $m'j$, notée $X_{y=m'_j}$.

Une distribution conditionnelle de X est une distribution particulière de X : c'est la liste des k modalités de X associées chacune à un effectif ou une fréquence ; elle est particulière en ce sens qu'elle ne concerne que les individus d'un sous-échantillon, et non tous les individus de E, comme dans le cas de la distribution margianle.

Il faut également remarquer les rôles différents et complémentaires des deux variables : X est la variable *conditionnée* dont on étudie les distributions, Y est la variable *conditionnant*, servant à construire les sous-échantillons sur lesquels on veut comparer X.

De manière symétrique, on définit une distribution conditionnelle de Y comme une distribution de Y restreinte à un des sous-échantillon conditionné par une modalité de X ; il y en donc k : par exemple, la distribution de Y sur le sous-échantillon E_{m_i} est la distribution de Y conditionnée par m_i, notée Y_{m_i}, ou encore la distribution de Y sachant que X vaut m_i, notée $Y_{x=m_i}$. Dans ce cas, X est la variable conditionnant et Y la variable conditionnée.

6 déf Considérons la distribution conditionnelle $X_{m'_j}$ (la distribution de X conditionnée par la modalité m'_j de Y) ; c'est la distribution de X restreinte aux $n_{.j}$ individus du sous-échantillon $E_{m'_j}$: ses effectifs sont donc les nombres de la jème *colonne* du tableau de contingence *en effectif*, $n_{1j}, n_{2j}, ..., n_{kj}$; le tableau suivant reprend en ligne cette jème colonne :

X	m_1	m_2	...	m_i	...	m_k	$Total$
Effectif	n_{1j}	n_{2j}		n_{ij}		n_{kj}	$n_{.j}$

Exemple : distribution en effectif de X conditionnée par la modalité *Rare*, X_{Rare}

X	A	B	Total
Effectif	7	8	15

La fréquence de la modalité m_i de $X_{m'_j}$ est la fréquence de m_i dans le sous-échantillon $E_{m'_j}$ et non pas dans l'échantillon E tout entier ; on la calcule donc en divisant chaque effectif n_{ij} par la taille $n_{.j}$ du sous-échantillon, et non par n, comme dans la distribution conjoite *en fréquence*.

Exemple : distribution en fréquence de X_{Rare}

X	A	B	Total
Fréquence	7/15=0.47	8/15=0.53	1

7 À partir du tableau de contingence on peut donc construire $p+1$ distributions de X : p distributions conditionnelles, $X_{m'_1}, X_{m'_2}, ..., X_{m'_p}$, correspondant chacune à une colonne et la distribution marginale dite aussi globale :

X	$X_{m'_1}$	$X_{m'_2}$	...	$X_{m'_j}$	...	$X_{m'_p}$	*Globale*
m_1	n_{11} ou $n_{11}/n.1$	$n_{12}[/n_{.2}]$	...	$n_{1j}[/n_{.j}]$	...	$n_{1p}[/n_{.p}]$	$n_{1.}[/n]$
m_2	n_{21} ou $n_{21}/n.1$	$n_{22}[/n_{.2}]$	...	$n_{2j}[/n_{.j}]$	...	$n_{2p}[/n_{.p}]$	$n_{2.}[/n]$
...	...	...	...	...	...	...	...
m_i	n_{i1} ou $n_{i1}/n.1$	$n_{i2}[/n_{.2}]$	...	$n_{ij}[/n_{.j}]$	...	$n_{ip}[/n_{.p}]$	$n_{i.}[/n]$
...	...	...	...	...	...	...	...
m_k	n_{k1} ou $n_{k1}/n.1$	$n_{k2}[/n_{.2}]$	...	$n_{kj}[/n_{.j}]$	...	$n_{kp}[/n_{.p}]$	$n_{k.}[/n]$
Total	$n_{.1}$ ou 1	$n_{.2}$ ou 1	...	$n_{.j}$ ou 1	...	$n_{.p}$ ou 1	n ou 1

De manière équivalente, on peu construire $k+1$ distributions de Y : k distributions conditionnelles, $Y_{m_1}, Y_{m_2}, ..., Y_{m_k}$, correspondant chacune à une ligne, et la distribution marginale (ou globale) :

Y	m'_1	m'_2	...	m'_j	...	m'_p	*Total*
Y_{m_1}	n_{11} ou $n_{11}/n_{1.}$	n_{12} ou $n_{12}/n_{1.}$	...	n_{1j} ou $n_{1i}/n_{1.}$	...	n_{1p} ou $n_{1p}/n_{1.}$	$n_{1.}$ ou 1
Y_{m_2}	$n_{21}[/n_{2.}]$	$n_{22}[/n_{2.}]$	...	$n_{2j}[/n_{2.}]$	...	$n_{2p}[/n_{2.}]$	$n_{2.}$ ou 1
...	...	...	...	...	...	...	...
Y_{m_i}	$n_{i1}[/n_{i.}]$	$n_{i2}[/n_{i.}]$	...	$n_{ij}[/n_{i.}]$	...	$n_{ip}[/n_{i.}]$	$n_{i.}$ ou 1
...	...	...	...	...	...	...	...
Y_{m_k}	$n_{k1}[/n_{k.}]$	$n_{k2}[/n_{k.}]$	...	$n_{kj}[/n_{k.}]$	...	$n_{kp}[/n_{k.}]$	$n_{k.}$ ou 1
Globale	$n_{.1}[/n]$	$n_{.2}[/n]$	...	$n_{.j}[/n]$	...	$n_{.p}[/n]$	n ou 1

8 **Exemple « niveau scolaire et absentéisme ».** Distributions conditionnelles de X, le niveau scolaire, en effectif et en (fréquence) : il y en a 3, comme le nombre de modalités de Y, disposées verticalement comme le sont les 2 modalités de X :

Mod X	X_{Rare}	X_{Moyen}	$X_{Frequent}$	X Global
A	7 (0,47)	4 (0,67)	4 (0,67)	15 (O,56)
B	8 (0,53)	2 (0,33)	2 (0,33)	12 (O,44)
Total	15 (1)	6 (1)	6 (1)	27

Distributions conditionnelles de Y, l'absentéisme, en effectif : il y en a 2, comme le nombre de modalités de X, disposées horizontalement comme le sont les 3 modalités de Y :

mod. Y	Rare	Moyen	Fréquent	Total
Y_A	7	4	4	15
Y_B	8	2	2	12
Y globale	15	6	6	27

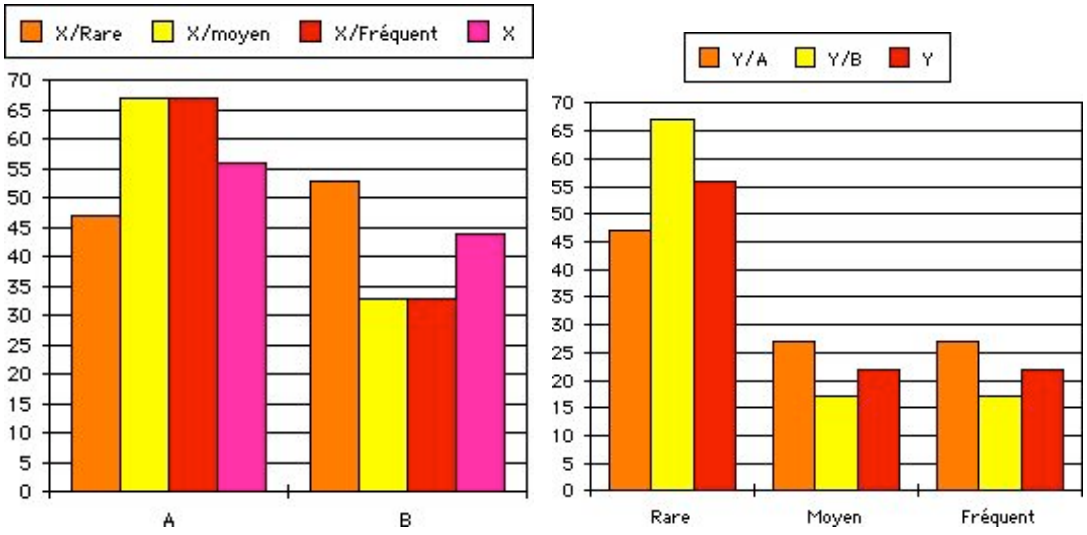
Représentations graphiques

Représentation simultanée des conditionnelles

9 Si la variable n'est pas continue, on peut représenter sur le même graphique les diagrammes en barre de ses conditionnelles :

1. on place horizontalement les modalités de la variable, dans l'ordre si elle est ordonnée, sur un axe unitaire si elle est numérique ;
2. au-dessus de chaque modalité, on trace une barre dont la longueur est égale ou proportionnelle à son effectif.

Exemple : représentation simultanée des distributions conditionnelles en fréquence du niveau scolaire et de l'absentéisme (on a rajouté les distributions marginales) :

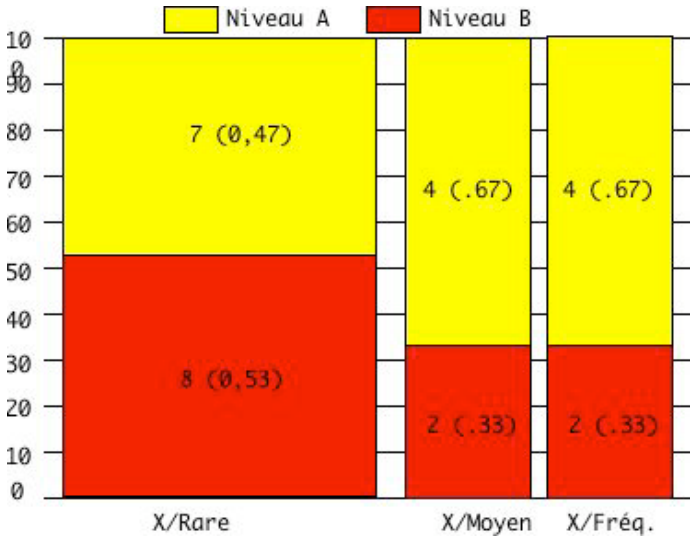


Représentation de la distribution conjointe à l'aide des conditionnelles

10 La représentation de la distribution conjointe de X et Y consiste à dessiner les distributions conditionnelles d'une des variables dans les sous-échantillons induits par l'autre ; on peut le faire de deux manières : dessiner les conditionnelles de X dans les sous-échantillons induits par Y, ou dessiner les conditionnelles de Y dans les sous-échantillons induits par X ; détaillons la première manière :

1. on représente les sous-échantillons induits par Y dans des rectangles de hauteurs identiques et de largeur égale à leur taille : ainsi, en convenant que les hauteurs égales à, les surfaces sont égales aux tailles ;
2. on divise ensuite chaque sous-échantillon en rectangles de hauteur égale aux fréquences de la distribution conditionnelle de X associée.

Exemple : pour représenter la distribution conjointe avec les conditionnelles du niveau scolaire X, on dessine les trois rectangles représentant les sous-échantillons induits par l'absentéisme Y, de surface égale à leur effectif 15, 6 et 6 ; ensuite, on divise chaque rectangle selon la distribution conditionnelle de X associée, X_{Rare}, X_{Moyen} et X_{Freq} :



11 Caractéristiques de la représentation.

- Chaque division a une surface égale à l'effectif correspondant du tableau de contingence : par exemple, le rectangle en haut et à gauche a une largeur de $n_{.1}$ et une hauteur de $n_{11}/n_{.1}$, et donc une surface égale à n_{11}.
- Les surfaces sont des effectifs : une plus grande surface est l'indication d'un effectif plus important.
- Les hauteurs sont des fréquences : une plus grande hauteur est l'indication d'une fréquence plus importante.
- On compare les surfaces ou les hauteurs selon qu'on veut faire une comparaison en valeur absolue ou en valeur relative.

Exemple : le rectangle supérieur gauche a une surface de 15*0,47=7 égale à l'effectif n_{11} de la modalité conjointe $(A, Rare)$ auquel il est associé ; le rectangle supérieur droit à une surface égale à l'effectif n_{13} de la modalité conjointe $(A, Frequent)$ crrespondante.

En comparant les surfaces de ces deux rectangle on peut affirmer que les élèves de niveau A sont *plus* nombreux parmi les élèves rarement absents que parmi les élèves fréquemment absents : 7 contre 4 ; par contre, en comparant leur hauteur, on peut affirmer qu'ils sont *relativement moins* nombreux parmi les élèves rarement absents que parmi les élèves fréquemment absents : 47% contre 67%.

Questions de cours

1. Qu'appelle-t-on sous-échantillon induit par la modalité m'_4 de Y ? Quelles sont ses deux notations ?
2. E_{m_4} est-elle la notation simplifiée de $E_{x=m_4}$ ou de $E_{y=m_4}$?
3. Combien y-a-t-il de sous-échantillons induits par X ?
4. Quelle est la taille des sous-échantillons induits E_{m_4} et $E_{m'_2}$
5. Qu'appelle-t-on partition de E conditionnée par Y ?
6. Dans quel contexte peut-on parler de distribution conditionnelle ?
7. Définition d'une distribution conditionnelle ?
8. Combien y-a-t-il de distributions conditionnelles de Y ?
9. Combien y-a-t-il de modalités dans une distribution conditionnelle de Y ?
10. Les notations suivantes ont-elles un sens : $X_{m'_4}$, X_{m_4}, $Y_{m'_4}$, Y_{m_4} ?
11. À partir d'un tableau de contingence, combien peut-on produire de distribution de X ? de Y ?
12. Sur un tableau de contingence en effectif comment lit-on les distributions conditionnelles de X en effectif ?
13. Sur un tableau de contingence en fréquence comment lit-on les distributions conditionnelles de X en fréquence ?
14. Quelle est la fréquence de la seconde modalité de la distribution conditionnelle $X_{m'4}$? Y_{m3} ?
15. Quelle sont les effectifs totaux des distributions conditionnelles de X ? À quoi est égale leur total ?

Questions sur le cours

1. Pourquoi ne fait-on (généralement) pas de représentation simultanée de conditionnelles continues ?
2. Vérifier que la représentation de la distribution conjointe de l'exemple à l'aide des conditionnelles de l'absentéisme est de la forme :