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

Cours 4
Liaison entre deux variables conjointes

1 déf **La liaison statistique entre deux variables X et Y** exprime l'information que donne la mesure de l'une des deux variables pour un individu de la population, sur la valeur de l'autre variable pour ce même individu : par exemple, la liaison de Y à X est l'information que donne la mesure x_e de X effectuée sur un individu e, pour estimer la valeur y_e, cette indication étant d'autant plus précise que la liaison est forte.

L'étude d'une liaison statistique répond à deux types de problématiques :
- sur le plan de la connaissance, elle permet de suggérer une éventuelle relation de causalité entre les deux variables ; et si elle ne peut l'expliquer (voir plus loin), elle encourage la recherche de son explication ;
- sur le plan pratique, elle permet d'avoir une évaluation de la mesure d'une des deux variables, disons y, à partir de la seule mesure de l'autre, x ; ceci est particulièrement utile lorsque la mesure de la première s'avère être difficile, très coûteuse, voire impossible.

Dans ce cours nous étudierons la liaison statistique entre les deux variables à partir de leur observation conjointe sur un échantillon ; nous supposerons que les conclusions auxquelles nous parvenons peuvent se généraliser à la population toute entière, ce qui est une autre façon de dire que l'échantillon est « représentatif » de la population.

2 La liaison statistique se manifeste principalement sur les distributions conditionnelles ; supposons par exemple que nous étudions la liaison de Y à X ; pour une mesure $x = m_i$ de X effectuée sur un individu de la population, la distribution conditionnelle Y_{m_i} qui est la distribution des valeurs de y dans le *sous-échantillon* E_{m_i}, est sensiblement identique à la distribution de Y dans la *sous-population* P_{m_i} des individus de P dont la valeur x est m_i ; Y_{m_i} est donc la distribution des valeurs possibles de y quand $x = m_i$.

Prenons comme illustration l'exemple suivant dans lequel les distributions conditionnelles de Y sont données en pourcentage :

X \ Y	m'_1	m'_2	m'_3	m'_4	Total
m_1	15	50	30	5	100
m_2	70	15	10	5	100
m_3	5	10	35	50	100

- la première distribution conditionnelle Y_{m_1} est une estimation de la distribution de Y sur la sous-population P_{m_1} des individus ayant m_1 comme mesure pour X (l'échantillon étant supposé « représentatif ») ; si la mesure x d'un individu de la population est la valeur m_1, il y a de fortes chances que sa mesure y soit m'_2 ou m'_3 puisque 80% (environ) des individus de P_{m_1} ont l'une de ces deux modalités comme mesure pour Y ;
- de la même façon, si $x = m_2$, la valeur la plus probable pour y est m'_1 ;
- enfin si $x = m_3$ la valeur de de y risque d'être m'_4 ou m'_3.

3 **Information nulle et indépendance.** Dans le cas limite où les distributions conditionnelles de Y sont égales, le fait de savoir qu'un individu a la valeur x pour X n'a aucune utilité pour estimer sa mesure pour Y : l'information apportée par la mesure de X est nulle : on dit que `Y est indépendant de X`.

De manière analogue, si les distributions conditionnelles de X sont égales, l'information donnée par y concernant x est nulle : X est indépendant de Y.

L'indépendance est étudiée en détail dans les chapitres suivants ; nous y verrons que ce cas extrême d'égalité des distributions conditionnelles dans un échantillon ne se rencontre malheureusement jamais en pratique, et qu'il faudra mettre en œuvre un outillage statistique plus élaboré, des tests, pour suggérer l'indépendance des deux variables.

4 Information totale et liaison fonctionnelle. Le cas limite inverse est celui où l'information donnée par x permet de déterminer sans ambiguïté une valeur et une seule y ; l'information apportée est totale, et on dit que `Y est liée fonctionnellement à X` ; c'est souvent une fonction f qui lie Y à X : la valeur y est alors calculée à partir de x par la formule $f(x)$.

Si Y est liée fonctionnellement à X, les distributions conditionnelles de Y ont une modalité et une seule de fréquence non nulle (donc égale à 1), ou, ce qui revient au même, chaque ligne du tableau de contingence contient un seul effectif non nul : si m'_j est la modalité de fréquence 1 dans la distribution conditionnelle Y_{m_i}, elle est la mesure de Y qu'on attribuera à un individu dont la mesure pour X est m_i. Dans l'illustration suivante, le tableau de contingence lie fonctionnellement Y à X, en indiquant la valeur de Y associée à chaque valeur de X :

X \ Y	m'_1	m'_2	m'_3
m_1	87	0	0
m_2	0	58	0
m_3	0	0	61
m_4	0	74	0

- si $x = m_1$, y vaut m'_4 ;
- si $x = m_2$, y vaut m'_2 ;
- si $x = m_3$, y vaut m'_3 ;
- si $x = m_4$, y vaut m'_2 ;

On remarquera que si ce tableau de contingence lie fonctionnellement Y à X, il ne lie pas fonctionnellement X à Y : si $y = m'_2$ le tableau indique deux valeurs pour x, m_2 et m_4, alors qu'il devrait n'en indiquer qu'une seule pour une liaison fonctionnelle.

5 Covariation. Quand les deux variables sont ordonnées, la liaison peut prendre la forme d'une covariation :

- covariation positive quand les valeurs de X et Y croissent ou décroissent « globalement » en même temps ; dans ce cas, le plus gros des effectifs du tableau de contingence se trouve sur la « diagonale » haut à gauche - bas à droite ;
- covariation négative quand les valeurs de X et Y croissent ou décroissent « globalement » en sens contraire ; le plus gros des effectifs se trouve alors sur la « diagonale » haut à droite - bas à gauche.

6 Remarque. La distribution conditionnelle Y_x est la meilleure information pour estimer y à partir de la valeur mesurée x. S'il fallait estimer cette valeur y en l'absence de la mesure x, la meilleure information serait la distribution marginale de Y, par le biais d'un de ses indices de localisation : moyenne, médiane, mode, etc.

7 Liaison statistique et causalité . La liaison statistique traite des relations formelles entre les variables X et Y, dans le modèle statistique de la situation, indépendamment donc de toute signification de cette situation ; une liaison statistique s'observe, se calcule, met en valeur des relations entre les valeurs observées ; mais elle n'explique rien, ne prouve rien : elle est seulement une indication, le signe d'une possible relation de causalité.

La causalité concerne les relations de nécessité entre les deux caractères modélisés par les variables ; elle s'étudie, se prouve par des méthodes propres au domaine de connaissances de ces caractères.

L'observation d'une liaison entre deux variables dans un échantillon ne permet pas de conclure à une relation de cause à effet entre les caractères modélisés ; cette liaison peut être une coïncidence comme l'évolution du prix des carburants et de la population de Montpellier par année ; elle peut également trouver son origine dans une cause commune, comme le coût des études et la profession dans une population d'anciens étudiants : c'est la cause commune qui en relation de cause à effet avec chaque caractère et non les deux caractères entre eux.

Rappels d'arithmétique

L'étude et l'analyse des distributions conditionnelles utilise les deux résultats suivants.

8 **Rappel 1 - Produit des extrêmes et des moyens dans une égalité de fractions.** Si deux fractions $\frac{a}{b}$ et $\frac{c}{d}$ sont égales, alors le produit des « extrêmes » $a * d$ est égal au produit des « moyens » $c * d$; et réciproquement.

Éléments de preuve : si $\frac{a}{b} = \frac{c}{d}$, on obtient le résultat en réduisant au même dénominateur : $\frac{a*d}{b*d} = \frac{c*b}{d*b}$ et donc $a * d = b * c$. Réciproquement, si $a * d = c * b$ on obtient le résultat en divisant les deux termes de l'égalité par $b * d$.

9 **Rappel 2 - Égalité de fractions.** Si des fractions sont égales, elles sont égales à la fraction obtenue en sommant numérateurs et dénominateurs :

$$\frac{n_1}{q_1} = \frac{n_2}{q_2} = ... = \frac{n_p}{q_p} = \frac{n_1 + ... + n_p}{q_1 + ... + q_p}$$

Éléments de preuve : on utilise le résultat précédant ; pour que $\frac{n_1}{q_1} = \frac{n_1+...+n_p}{q_1+...+q_p}$ il suffit que $n_1 * (q_1 + ... + q_p) = q_1 * (n_1 + ... + n_p)$; on le vérifie sans difficulté en remarquant les égalités $n_1 * q_2 = q_1 * n_2, n_1 * q_3 = q_1 * n_3, ..., n_1 * q_p = q_1 * n_p$

Propriétés des distributions conditionnelles

10 **Propriété 1.** Si les distributions conditionnelles de X en fréquence sont égales, alors elles sont égales à la distribution marginale de X en fréquence.

Éléments de preuve : sous l'hypothèse, les fréquences des modalités m_i sont égales entre elles : $\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = ... = \frac{n_{ip}}{n_{.p}}$; en appliquant le second rappel, chaque fraction est égale à $\frac{n_{i1}+...+n_{ip}}{n_{.1}+...+n_{.p}}$ qui vaut $\frac{n_{i.}}{n} = f_{i.}$, la fréquence marginale de m_i.

11 **Propriété 2.** Si les distributions conditionnelles de X (ou de Y) sont égales en fréquence, elles sont proportionnelles en effectif. Et réciproquement.

Éléments de preuve : les effectifs des deux distributions X_j et $X_{j'}$ sont dans le rapport des tailles des sous-échantillons, $\frac{n_{.j}}{n_{.j'}}$; en effet, puisque $\frac{n_{ij}}{n_{.j}} = \frac{n_{ij'}}{n_{.j'}}, n_{ij} = \frac{n_{.j}}{n_{.j'}} * n_{ij'}$.

12 **Propriété 3.** Si les distributions conditionnelles de X en fréquence sont égales alors on a l'égalité : $n_{ij} = n_{i.} * n_{.j}/n$ pour tous les i et j. Et réciproquement.

Éléments de preuve : c'est une conséquence directe de la propriété 1 et du premier rappel : si $\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$ alors $n_{ij} * n = n_{.j} * n_{i.}$ qui équivaut à $n_{ij} = \frac{n_{i.}*n_{.j}}{n}$

13 **Propriété 4.** Si les distributions conditionnelles de X en fréquence sont égales, alors les distributions conditionnelles de Y en fréquence le sont également.

Éléments de preuve : c'est une conséquence directe de propriété précédente.

Questions de cours

1. À quoi reconnaît-on que Y est indépendant de X sur le tableau de contingence ?
2. À quoi reconnaît-on que Y est liée fonctionnellement à X sur le tableau de contingence ?
3. X a trois modalités, Y quatre : construire un tableau de contingence en effectif qui montre l'indépendance de X par rapport à Y, avec des sous-échantillons induits par X et Y de tailles différentes.
4. X a trois modalités, Y quatre : construire un tableau de contingence qui montre une liaison fonctionnelle de X à Y.
5. Qu'en est-il des distributions conditionnelles si les lignes du tableau de contingence en effectif sont proportionnelles ?
6. Quelle est la valeur de n_{25} si les distributions conditionnelles de X en fréquence sont égales ?
7. X a trois modalités, Y quatre : construire un tableau de contingence en effectif pour lequel les distributions conditionnelles de X en fréquence sont égales, tandis que celles de Y ne le sont pas.

Questions sur le cours

1. En utilisant la propriété 3, déterminer si le tableau de contingence suivant suggère l'indépendance ou non de X par rapport à Y, de Y par rapport à X :

X \ Y	m'_1	m'_2	m'_3
m_1	36	189	75
m_2	84	441	175
m_3	11	65	24

2. Vérifier la propriété du produit des extrêmes et des moyens avec deux fractions à déterminer, puis avec la fraction $\frac{252}{715}$ et une autre à déterminer.