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

Cours 5
Indépendance

1 Dans le cours précédent, nous avons vu que la variable Y était indépendante de la variable X si ses distributions conditionnelles en fréquence sont égales ; dans ce cas en effet, la mesure de X sur un individu quelconque n'apporte pas d'information pour estimer la mesure de Y.

De la même manière X est indépendante de Y si ses distributions conditionnelles en fréquence sont égales.

À priori, on pourrait imaginer qu'une variable est indépendante de l'autre (par exemple X est indépendant de Y) sans que l'inverse soit vrai (Y n'est pas indépendant de X) ; mais les nombres utilisés dans les distributions sont ainsi faits que cette hypothèse est irréalisable.

Quelques propriétés des fractions et des distributions

2 **Produit des extrêmes et des moyens :** 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 ; la preuve est ici [1].

3 **Égalité de fractions :** si des fractions sont égales, elles sont égales à la fraction obtenue en sommant numérateurs et dénominateurs : $\frac{n_1}{q_1} = \frac{n_2}{q_2} = ... = \frac{n_p}{q_p} = \frac{n_1+...+n_p}{q_1+...+q_p}$; la preuve est ici [2].

4 **Égalité des conditionnelles et de la marginale :** si les distributions conditionnelles de X en fréquence sont égales, alors elles sont égales à la distribution marginale de X en fréquence ; la preuve est ici [3].

5 **Égalité des conditionnelles de X et de Y :** si les distributions conditionnelles de X en fréquence sont égales, alors les distributions conditionnelles de Y en fréquence le sont également ; et inversement ; la preuve est ici [4].

Indépendance des variables X et Y

Cette dernière propriété fait de la relation d'indépendance une relation symétrique ; elle énonce en effet que si les distributions conditionnelles *en fréquence* de l'une des deux variables sont égales, les distributions conditionnelles *en fréquence* de l'autre le sont également ; autrement dit, de deux choses l'une : ou bien les deux variables sont simultanément indépendantes de l'autre, ou bien aucune ne l'est ; on peut donc parler de l'indépendance des deux variables, plutôt que de l'indépendance de l'une par rapport à l'autre :

---

1. 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 $b * d$.

2. 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 $n_1 * q_2 = q_1 * n_2, n_1 * q_3 = q_1 * n_3, ..., n_1 * q_p = q_1 * n_p$

3. Si les distributions conditionnelles de X en fréquence sont égales, les colonnes du tableau de contingence en fréquence sont égales : pour chaque ligne i on a l'égalité des fractions : $\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = ... = \frac{n_{ip}}{n_{.p}}$; en appliquant la proproété du §3 , ces fractions sont égales à $\frac{n_{i1}+...+n_{ip}}{n_{.1}+...+n_{.p}}$ qui vaut $\frac{n_{i.}}{n} = f_{i.}$, la fréquence marginale de m_i.

4. Si les conditionnelles de X en fréquence sont égales alors (§4) on a l'égalité $\frac{n_{ij}}{n_{.j}} = f_{.j} = \frac{n_{i.}}{n}$ et donc l'égalité $n_{ij} = \frac{n_{i.}*n_{.j}}{n}$ (*), pour tous les effectifs ; en divisant les effectifs de la ième ligne par son total $n_{i.}$ on trouve que $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} = f_{.j}$ pour chaque colonne, ce qui signifie que la conditionnelle en fréquence Y_i est égale à la marginale de Y en fréquence, quelque soit i.

6 déf **Les variables X et Y sont dites indépendantes** si les distributions conditionnelles *en fréquence* de X et de Y sont *égales*.

On peut sans difficulté montrer que cette définition est équivalente à la suivante qui porte sur les distributions en effectifs (les lignes et les colonnes du tableau de contingence) : les distributions conditionnelles de X et de Y *en effectif* sont *proportionnelles*[5] .

7 Indépendance, population, échantillon. Les définitions précédentes valent aussi bien pour les distributions dans la population que dans l'échantillon : dans le premier cas on peut parler d'indépendance dans la population, dans le second d'indépendance dans l'échantillon.

Lorsque la problématique concerne l'indépendance dans la population et que les distributions conditionnelles dans la population ne sont pas évaluables, nous sommes contraints de faire l'étude de l'indépendance dans un échantillon, en supposant celui-ci « représentatif » pour pouvoir étendre les conclusions à la population toute entière. La question de la composition d'un échantillon « représentatif » est trop délicate pour que nous l'abordions ici : une branche de la discipline de la statistique y est consacrée, l'échantillonnage ; précisons seulement que supposer l'échantillon « représentatif » revient à considérer les distributions marginales du tableau de contingence comme « identiques » aux distributions de X et Y dans la population.

Distribution théorique d'indépendance

8 déf **Effectif théorique d'indépendance d'une modalité** . Si les variables sont indépendantes, les distributions conditionnelles en fréquences sont égales, et on a alors (§4) l'égalité $n_{ij} = \frac{n_{i.} * n_{.j}}{n}$ pour toutes les modalités conjointes ij ; cette valeur ne dépend que des marges, supposées être les distributions de X et Y dans la population ; autrement dit, le nombre $\frac{n_{i.} * n_{.j}}{n}$, calculé à partir des seules distributions de X et Y, est l'effectif qu'on doit observer dans un échantillon de taille n pour la modalité conjointe ij quand les variables sont indépendantes : on l'appelle effectif théorique d'indépendance de la modalité ij (ou effectif d'indépendance, ou effectif théorique) ; et on la note $\tilde{n}_{ij}$ (« n tilde de i et j»).

9 déf Soit D une distribution conjointe observée sur un échantillon ; la distribution conjointe dont les effectifs sont les effectifs d'indépendance $\tilde{n}_{ij}$ s'appelle la distribution théorique d'indépendance de D ou encore la la distribution théorique de D ; on la note $\tilde{D}$. Cette distribution se construit à partir des seules marges de D, à l'aide de la formule $\tilde{n}_{ij} = \frac{n_{i.} * n_{.j}}{n}$. C'est la distribution qu'on doit observer pour un échantillon de taille n quand X et Y sont indépendantes, en supposant que les distributions de X et Y dans la population sont données par les marges.

Exemple : à partir des marges de la distribution conjointe « niveau scolaire et absentéisme » :

X / Y	Rare	Moyen	Fréquent	Total X
A	7	4	4	15
B	8	2	2	12
Total Y	15	6	6	27

on construit la distribution d'indépendance suivante (par exemple, $\tilde{n}_{11} = 15 * 15/27 = 8.33$)

X / Y	Rare	Moyen	Fréquent	Total X
A	8,33	3,33	3,33	15
B	6,66	2,66	2,66	12
Total Y	15	6	6	27

5. Considérons deux conditionnelles de X quelconques, X_j et $X_{j'}$; si elles sont égales en fréquence on a pour toutes les modalités i $\frac{n_{ij}}{n_{.j}} = \frac{n_{ij'}}{n_{.j'}}$, et donc $n_{ij} = a * n_{ij'}$ avec $a = \frac{n_{.j}}{n_{.j'}}$: X_j et $X_{j'}$ en effectif sont proportionnelles, dans un rapport égal au rapport des tailles.

Cela signifie que si on suppose les distributions en proportion du niveau scolaire et de l'absentéisme dans la population égales à

A	B	Total
15/27	12/27	1

et

Rare	Moyen	Fréquent	Total X
15/27	6/27	6/27	1

on devrait observer 8,33 individus de niveau A avec une absence Rare dans un échantillon de taille 27, si les deux variables sont indépendantes ; ce nombre décimal donné par la théorie n'est pas réaliste puisque les effectifs d'une observation sont nécessairement des entiers, d'où sa dénomination d'effectif théorique ; on voit par là les premières limites de la théorie puisqu'il semble très illusoire de pouvoir rencontrer des situations statistiques pour lesquelles la distribution théorique d'indépendance serait uniquement composé d'entiers, autrement dit, de pouvoir faire des observations qui prouvent l'indépendance de deux variables.

Autre exemple : distribution théorique d'indépendance de l'observation « traitements anti-termites » (entre parenthèses les effectifs réellement observés) :

X / Y	T1	T2	T3	Total X
Contaminé	30.7 (26)	30.7 (48)	30.6 (18)	92
Sain	169.3 (174)	169.3 (152)	169.4 (182)	508
Total Y	200	200	200	600

10 Remarques.

- Même si les effectifs théoriques $\tilde{n}_{ij}$ sont des nombres décimaux, les distributions observées D et théoriques $\tilde{D}$ ont les même marges [6].
- Les notions d'effectif et distribution théorique d'indépendance permettent de formuler une nouvelle définition de l'indépendance : X et Y sont dites indépendantes si les effectifs observés n_{ij} sont identiques aux effectifs d'indépendance $\tilde{n}_{ij} = \frac{n_{i.} * n_{.j}}{n}$, ou bien si la distribution observée D est identique à la distribution d'indépendance $\tilde{D}$.
  Cette troisième définition équivalente signifie que dans l'hypothèse de l'indépendance de X et Y, les effectifs observés peuvent se calculer à partir des distributions marginales seules, autrement dit des distributions de X et Y dans la population ; ce qui revient à dire qu'une observation séparée des variables X et de Y donnent la même information qu'une observation conjointe.

Mesure locale de liaison

11 déf **Le taux de liaison d'une modalité conjointe** ij mesure un écart entre l'effectif observé et l'effectif qu'on devrait observer sous l'hypothèse d'indépendance ; c'est la différence normalisée entre l'effectif observé et l'effectif théorique de la modalité, notée t_{ij} : $t_{ij} = \frac{n_{ij} - \tilde{n}_{ij}}{\sqrt{\tilde{n}_{ij}}}$.

La normalisation est nécessaire pour prendre en compte la relativité des différences : une différence de 10 entre 1000 et 1010 n'a pas la même signification (1% d'augmentation) que la même différence de 10 entre 20 et 30 (50% d'augmentation) ; on normalise par $\sqrt{\tilde{n}_{ij}}$ (et non par $\tilde{n}_{ij}$ par exemple) pour une raison qu'on pourra expliquer par la suite.

On observe trois types de liaison locale :

l'indépendance locale $t_{ij} \approx 0$: tout se passe pour la modalité ij comme si X et Y étaient indépendantes.

l'attraction locale $t_{ij} > 0$: on observe plus fréquemment la modalité dans l'échantillon que si X et Y étaient indépendantes.

la répulsion locale $t_{ij} < 0$: on observe moins fréquemment la modalité dans l'échantillon que si X et Y étaient indépendantes.

6. Par exemple, l'effectif marginale de m_i dans $\tilde{D}$ est $\tilde{n}_{i.} = \sum_{j=1}^{p} \tilde{n}_{ij} = \sum_{j=1}^{p} \frac{n_{i.} * n_{.j}}{n} = n_{i.} * \frac{\sum_{j=1}^{p} n_{.j}}{n} = n_{i.} * \frac{n}{n} = n_{i.}$ qui est l'effectif marginal de la modalité m_i dans D.

12 ⚠ Comme les distributions marginales de D et $\tilde{D}$ sont identiques, les taux de liaison sont forcés de « s'équilibrer » sur chaque ligne et sur chaque colonne : une attraction, excès relatif d'observations, doit s'accompagner d'une répulsion, défaut relatif d'observations, sur la même ligne et la même colonne ; même chose pour les répulsions ; aussi, l'observation de certaines attractions ou répulsions peut s'expliquer non comme une caractéristique de la liaison, mais comme un artefact, comme la conséquence mécanique d'une autre répulsion ou attraction sur la même ligne ou sur la même colonne.

13 Exemples :

1 – Niveau scolaire et absentéisme : le calcul de t_{11} donne $\frac{7-8.33}{\sqrt{8.33}} = -0.46$

X / Y	Rare	Moyen	Fréquent
A	-0,46	0,36	0,36
B	0,51	-0,4	-0,4

Le tableau met en évidence les attractions (A ; Moyen Fréquent) (B ; Rare), et les répulsions (A ; Rare) (B ; Moyen Fréquent) ; les lignes et les colonnes sont « équilibrées ».

2 – Traitements anti-termites :

X / Y	T1	T2	T3
Contaminé	-0,84	3,16	-2,29
Sain	0,36	-1,33	0,97

La forte attraction pour la modalité (Sain ; T3) pourrait exprimer l'efficacité du traitement T3 ; les autres valeurs seraient alors des « effets de bord » rendus nécessaires par l'équilibrage des lignes et des colonnes (sur la troisième colonne par exemple, il faut une forte répulsion pour compenser la forte attraction).

Mesure globale de liaison : la distance du χ^2

14 Dans la pratique, la distribution observée D n'est jamais identique à la distribution d'indépendance $\tilde{D}$, même quand on sait que X et Y sont indépendantes : la raison en est due aux fluctuations d'échantillonnage, l'objet du prochain cours ; si bien que pour étudier l'indépendance de X et Y, nous allons devoir juger non de *l'égalité* de D et $\tilde{D}$, mais de la *proximité* entre D et $\tilde{D}$.

15 déf **Distance du χ^2.** C'est une mesure de l'écart entre une distribution conjointe observée D et sa distribution théorique d'indépendance $\tilde{D}$; sa valeur est la somme des carrés des taux de liaisons :

$$\chi^2(D) = \sum_{1\leq i\leq k, 1\leq j\leq p} (t_{ij})^2 = \sum_{1\leq i\leq k, 1\leq j\leq p} \frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

- Le terme $t_{ij}^2 = \frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}$ est un nombre décimal positif ou nul ; on l'appelle contributions au χ^2 de la modalité conjointe ij ; $\chi^2(D)$ est donc composée de $k * p$ contributions.
- Si $D = \tilde{D}$ le nombre $\chi^2(D)$ est évidemment nul (chaque contribution est nulle).
- Si le nombre $\chi^2(D)$ est nul alors $D = \tilde{D}$: étant une somme de nombres positifs ou nuls, il ne peut s'annuler que si tous les termes sont nuls, autrement dit si les effectifs observés n_{ij} sont égaux aux effectifs théoriques $\tilde{n}_{ij}$.
- Ces deux dernières remarques suggèrent une nouvelle définition équivalente de l'indépendance : X et Y sont indépendantes si $\chi^2(D) = 0$.

16 Exemple :

1 – Niveau scolaire et absentéisme. La contribution de la modalité (A,Rare) est égale à $(7 - 8{,}33)^2/8{,}33 = 0{,}21$ (-0.46^2) ; le χ^2 de cette distribution conjointe est égal à 1,05.

X / Y	Rare	Moyen	Fréquent	Total
A	0,21	0,13	0,13	
B	0,26	0,16	0,16	
				1,05

2 – Traitements anti-termites. La contribution de la modalité (Contaminé,T1) est égale à $(26 - 30{,}7)^2/30{,}7 = 0{,}71$ (-0.84^2) ; le χ^2 de cette distribution conjointe est 18,59.

X / Y	T1	T2	T3	Total
Contaminé	0,71	9,8	5,23	
Sain	0,13	1,77	0,95	
Total				18,59

Questions de cours

1. Énoncer 3 définitions de l'indépendance de deux variables conjointes.
2. Quelle est la conséquence sur les distributions marginales, de supposer que l'échantillon est représentatif ?
3. Qu'appelle-t-on effectif d'indépendance ?
4. Que désigne $\tilde{n}_{ij}$?
5. Comment se calcule $\tilde{n}_{ij}$?
6. A-t-on besoin de la distribution conjointe pour calculer $\tilde{n}_{ij}$?
7. Définition de la distribution théorique d'indépendance ?
8. Comment construit-on la distribution théorique d'indépendance de D ?
9. Que signifie le taux de liaison t_{ij} ?
10. Quelle est la valeur de t_{34} ?
11. Comment reconnaît-on une indépendance locale ? Que signifie-t-elle ?
12. Comment reconnaît-on une attraction locale ? Que signifie-t-elle ?
13. Comment reconnaît-on une répulsion locale ? Que signifie-t-elle ?
14. Un taux de liaison positif marque-t-il nécessairement une attraction locale des deux variables ?
15. Que représente le χ^2 d'une distribution D ?
16. Donner une condition pour que $\chi^2(D)$ soit nul ?
17. Donner une condition pour que $\chi^2(D)$ soit négatif ?
18. Qu'appelle-t-on contribution d'une modalité au χ^2 ?
19. Combien y-a-t-il de contributions dans le calcul du χ^2 ?

Question sur le cours

1. Montrer que l'effectif marginal de m_i de D est identique à celui de $\tilde{D}$.
2. Montrer que $\chi^2(D) = 0$ si les distributions conditionnelles de X sont égales en fréquence.
3. Montrer que X et Y sont indépendantes si $\chi^2(D) = 0$.