Méthodes quantitatives des sciences sociales

4. Liaison, corrélation et causalité entre deux variables

Sciences Po Saint-Germain-en-Laye, 1ère année

2016-2017

Introduction

- La mise en relation de deux variables est un premier moment dans la recherche de **liens de causalité:** idée que A => B (ou B =>A ?). « Effet » d'une variable.
- On parle ici de **statistique bivariée**.
- Les variables peuvent être 1. toutes les deux **quantitatives** (numériques), 2. toutes les deux **qualitatives** (catégorisées), 3. ordinales ou encore 4. l'une quantitative l'autre qualitative. Dans chaque cas, on utilise une démarche particulière pour définir et calculer le **sens** et l'**intensité** du lien entre les deux variables.
- NB: On étudiera aujourd'hui les deux premiers cas (le troisième cas est spécifique, le dernier cas est traité grâce à l'**analyse de la variance** – ANOVA).

Plan

- 1. Le lien entre deux variables: de la visualisation à l'interprétation causale
- 2. Le coefficient de corrélation bivariée et la régression linéaire simple
- 3. La liaison entre deux variables qualitatives: les tableaux de contingence
- 4. De la description à l'inférence statistique

Bibliographie

- R.Boudon, *Les méthodes en sociologie*, Paris, PUF, 1969.
- H.Le Bras, E.Todd, *L'invention de la France. Atlas anthropologique et politique*, Paris, Gallimard, 2012.
- O.Martin, *L'analyse quantitative des données*, Paris, Armand Colin, 2012.
- J.-M.Meunier, *Statistiques pour psychologues*, Paris, Dunod, 2010.

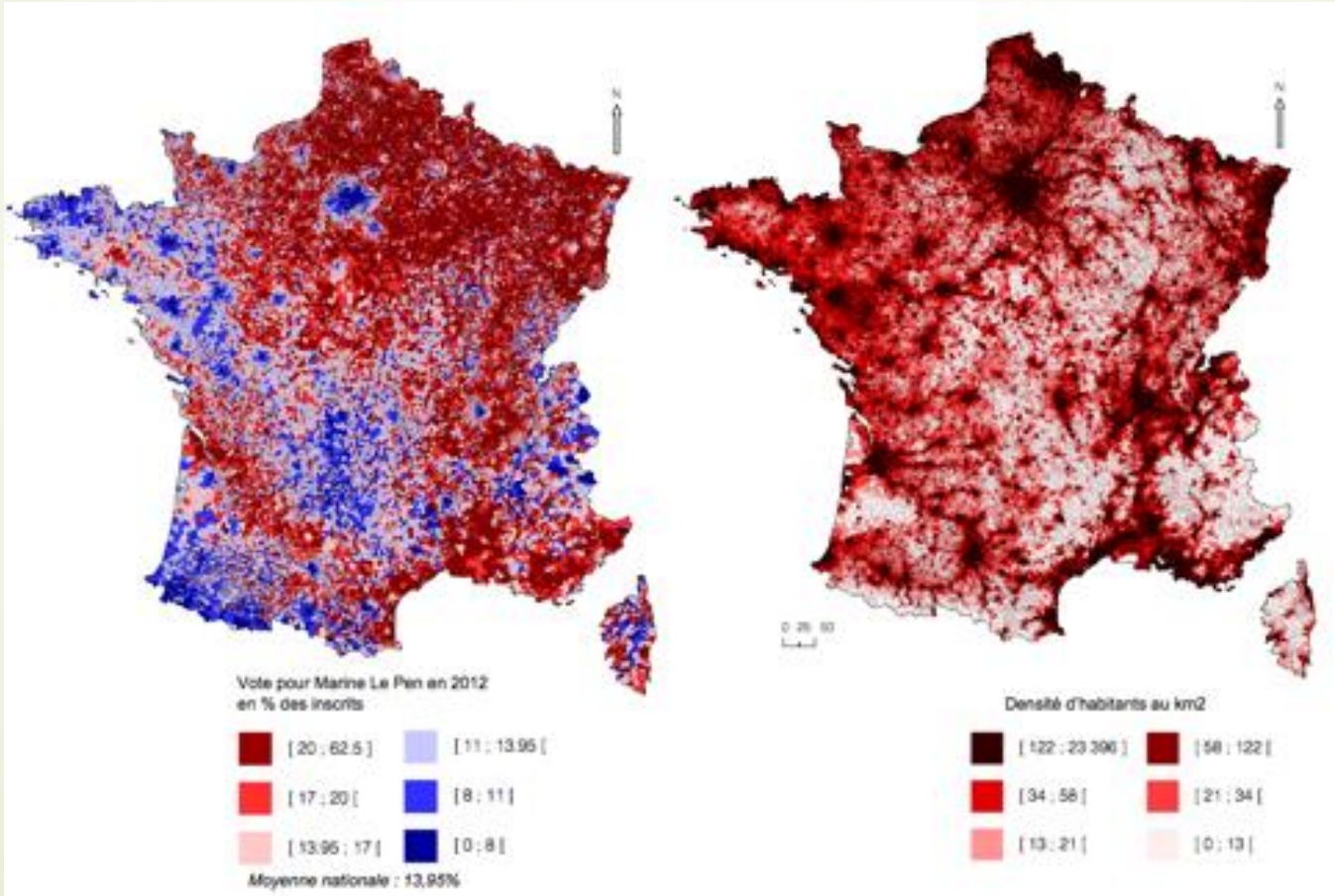
1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- Il existe plusieurs façons de mettre en évidence des corrélations ou des liaisons entre variables. Exemple déjà vu (chap. 1): Durkheim met en évidence une relation entre l'âge et la propension au suicide en comparant des fréquences par classes d'âge.
- On utilise des techniques de **visualisation** pour faire apparaître des **corrélations**.
- Cartographie, comparaison de séries chronologiques, diagrammes de dispersion: trois façons complémentaires de représenter la corrélation entre deux **variables numériques**.

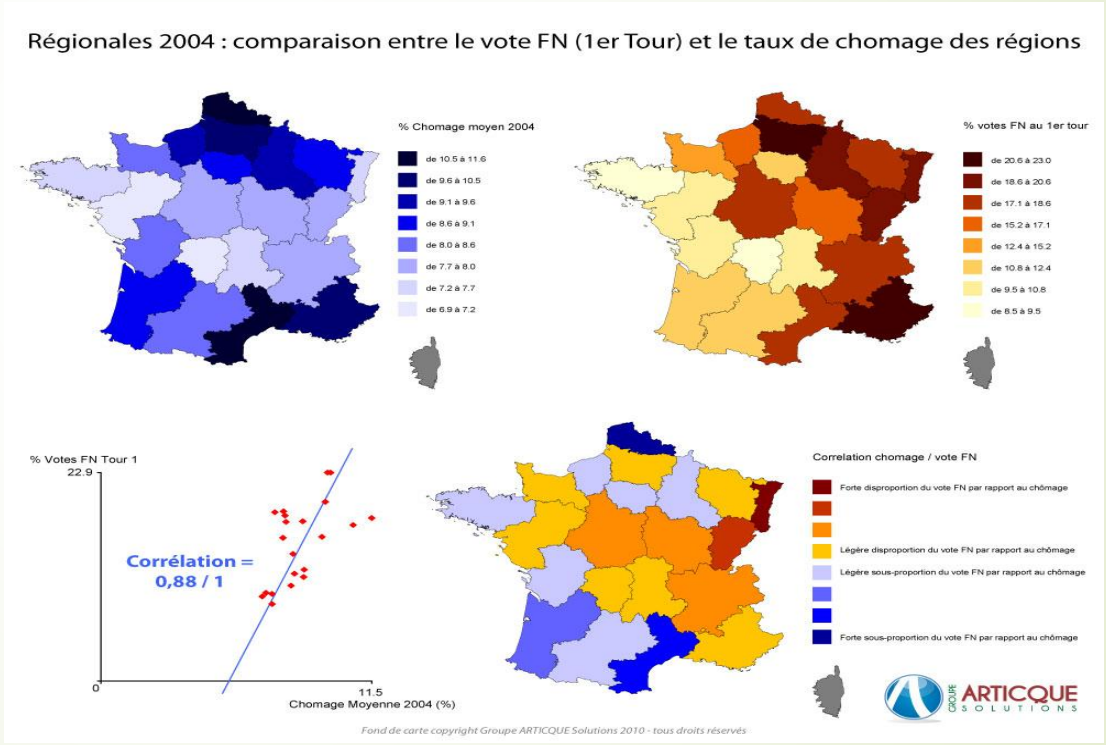
1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- La **cartographie**. Elle est abondamment utilisée par exemple par Hervé Le Bras et Emmanuel Todd dans *L'invention de la France. Atlas anthropologique et politique (2012).* Elle leur permet de mettre en relation les structures familiales, religieuses et les comportements, notamment politiques, en très longue période.
- **Démarche**: on juxtapose deux cartes après avoir colorié pour chacune d'elles les unités en fonction des valeurs prises par un indicateur (mis en classe, donc ordinal).
- Quand les cartes « se ressemblent », on met en évidence une corrélation. Exemple: un indicateur de vote (% de vote pour tel candidat, etc.) et un indicateur socio-économique ou socio-démographique.

1. Le lien entre deux variables: de la visualisation à l'interprétation causale



1. Le lien entre deux variables: de la visualisation à l'interprétation causale



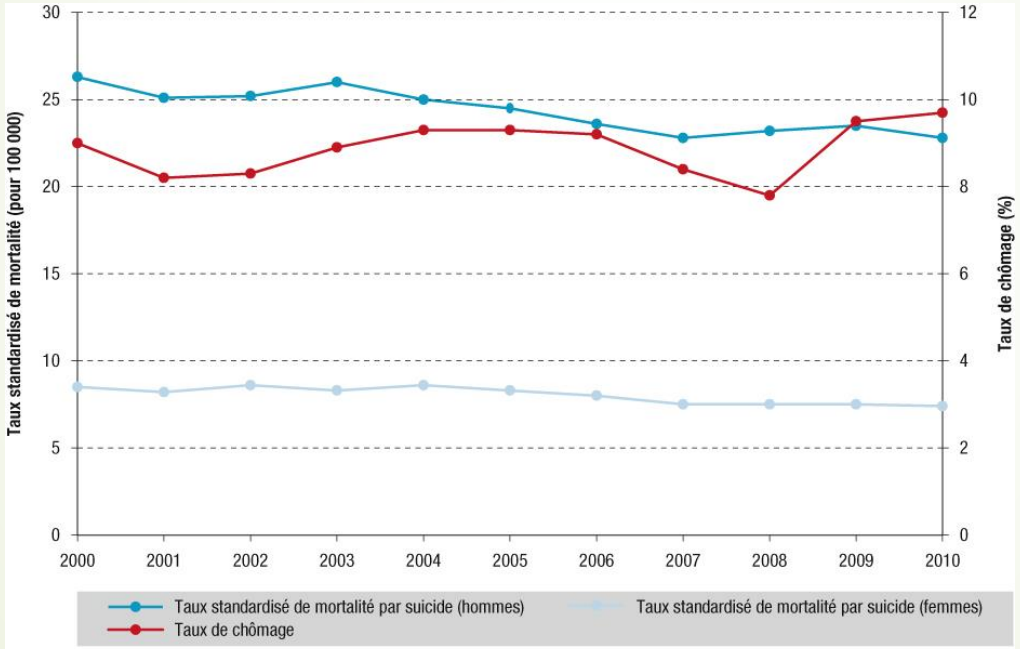
1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- Une pratique qui s'est généralisée avec la diffusion des logiciels de cartographie (ex: ArcGis…).
- Le problème le plus connu est le « paradoxe écologique »: une corrélation ainsi mesurée ne signifie pas nécessairement que la relation existe au niveau des individus considérés (exemple: pas de relation entre proportion de personnes âgées et vote Sarkozy dans un département, ***mais*** les données de sondage montrent plutôt une relation forte entre les deux variables. D'autres facteurs différencient les départements).
- Une question plus générale: nature du lien ainsi établi. L'interprétation causale suppose toujours un cadre ou un **schéma causal**. Dans le cas de Le Bras et Todd, celui-ci est anthropologico-religieux et géographique.

1. Le lien entre deux variables: de la visualisation à l'interprétation causale

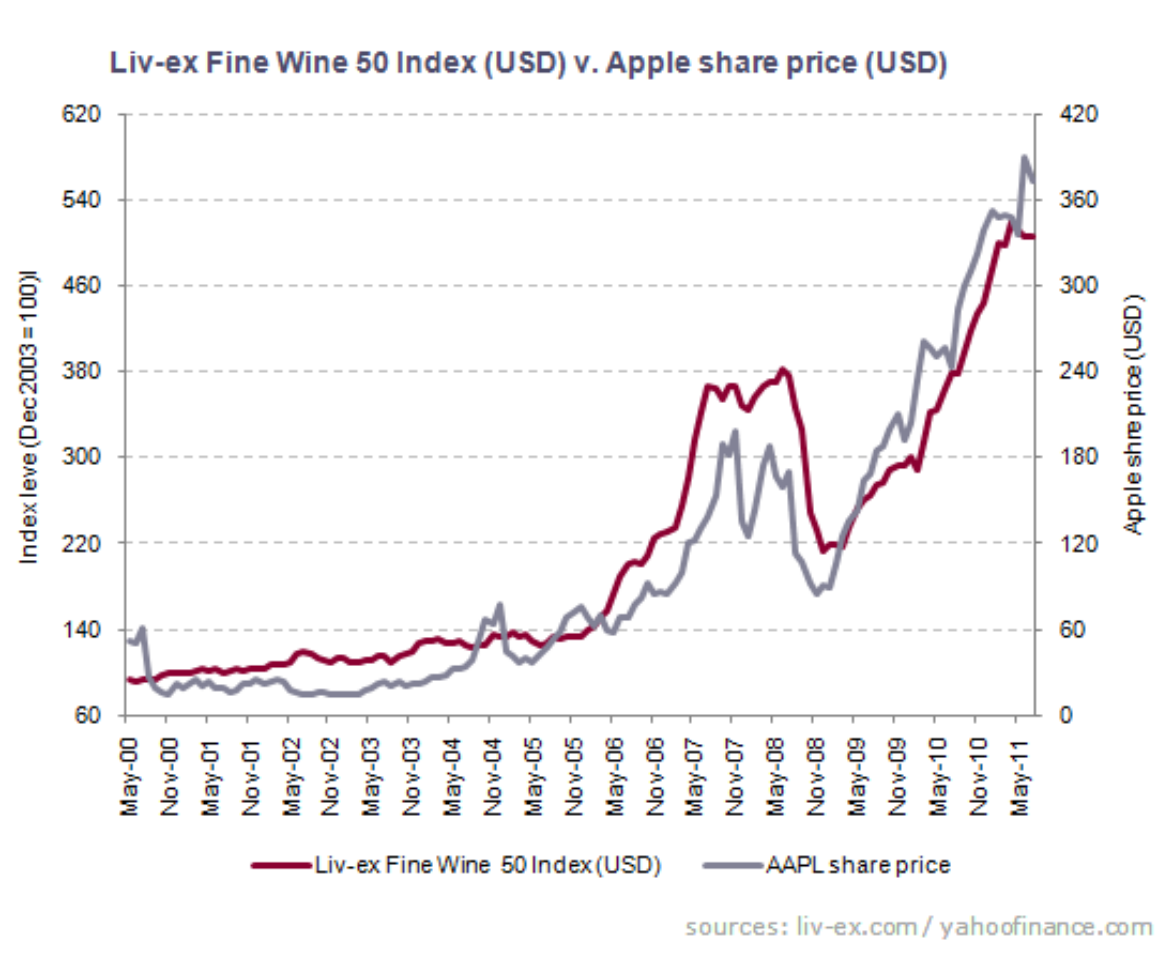
- Une autre démarche tout aussi classique consiste à figurer simultanément deux courbes représentant des **séries chronologiques**. Cette fois, on montre une relation entre deux variables non pas dans l'espace mais dans le temps.
- Pour chaque moment, on reporte la valeur des deux variables (en utilisant parfois deux échelles distinctes à droite et à gauche).
- C'est la démarche qu'employait François Simiand (*Le salaire, l'évolution sociale et la monnaie, 1932*) pour expliquer la survenue des grandes crises économiques ou cycles longs (« phase A » et « phase B » de l'économie). (Rôle de l'expansion monétaire => hausse des prix => hausse des salaires…).
- On l'utilise toujours en analyse de conjoncture, et dès que l'on dispose de suffisamment d'observations pour deux séries, on peut les représenter de cette façon.

1. Le lien entre deux variables: de la visualisation à l'interprétation



Taux de mortalité par suicide standardisés sur la population européenne, IARC 1976.

1. Le lien entre deux variables: de la visualisation à l'interprétation causale



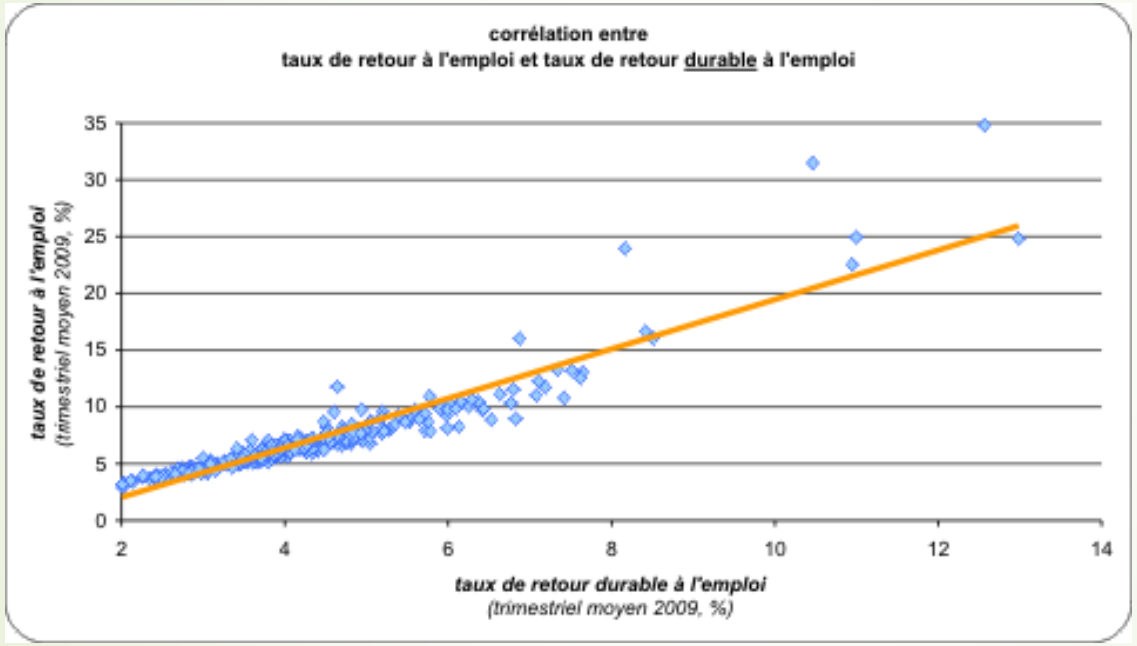
1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- On retrouve ici le problème plus général de la nature du lien ainsi établi. Exemple: les séries financières sont souvent fortement corrélées. Cela n'implique pas nécessairement de lien causal entre elles. A nouveau, c'est le schéma causal qui permet d'**interpréter** l'existence d'une corrélation entre deux variables.
- On est confronté ici à un autre problème, celui de l'existence de fluctuations, ou cycles, que les auteurs cherchent aussi à interpréter à l'aide d'un schéma causal.
- C'est le domaine de l'**économétrie des séries temporelles**, avec les discussions autour de la causalité (« Granger causality »).

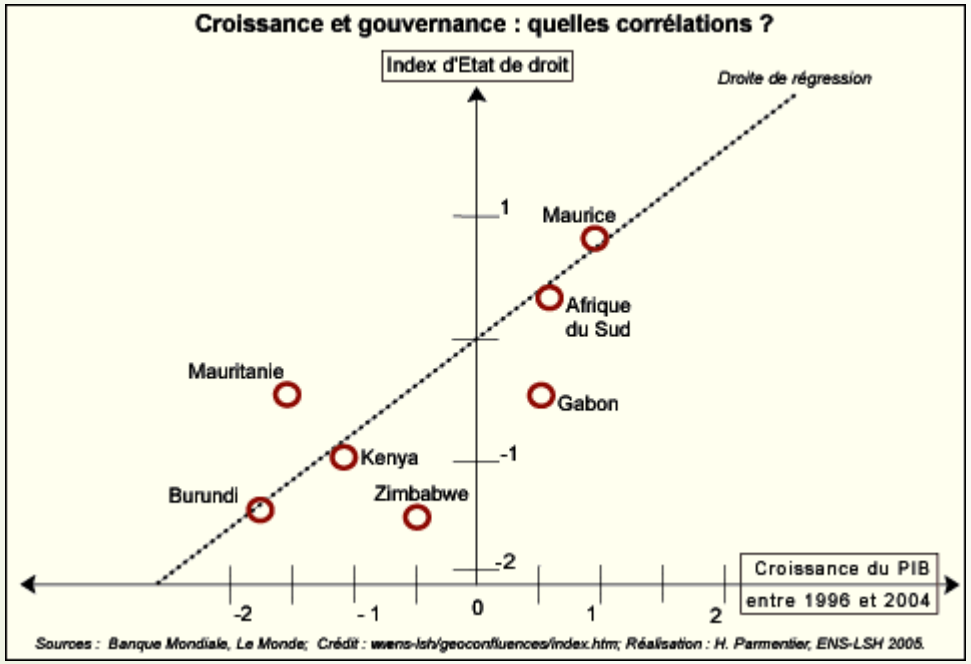
1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- Une troisième représentation classique de la relation entre deux variables numériques est le **diagramme de dispersion** ou diagramme de corrélation.
- On représente une variable en abscisse et l'autre en ordonnée et on associe un point à chaque couple de valeurs.
- Si la relation semble **linéaire**, on visualise la force de la corrélation, et aussi son sens, en figurant la **droite de régression**. A nouveau, on suggère fortement un lien… ou l'absence de lien (au sens de relation **linéaire**, mais on peut imaginer toute relation fonctionnelle).

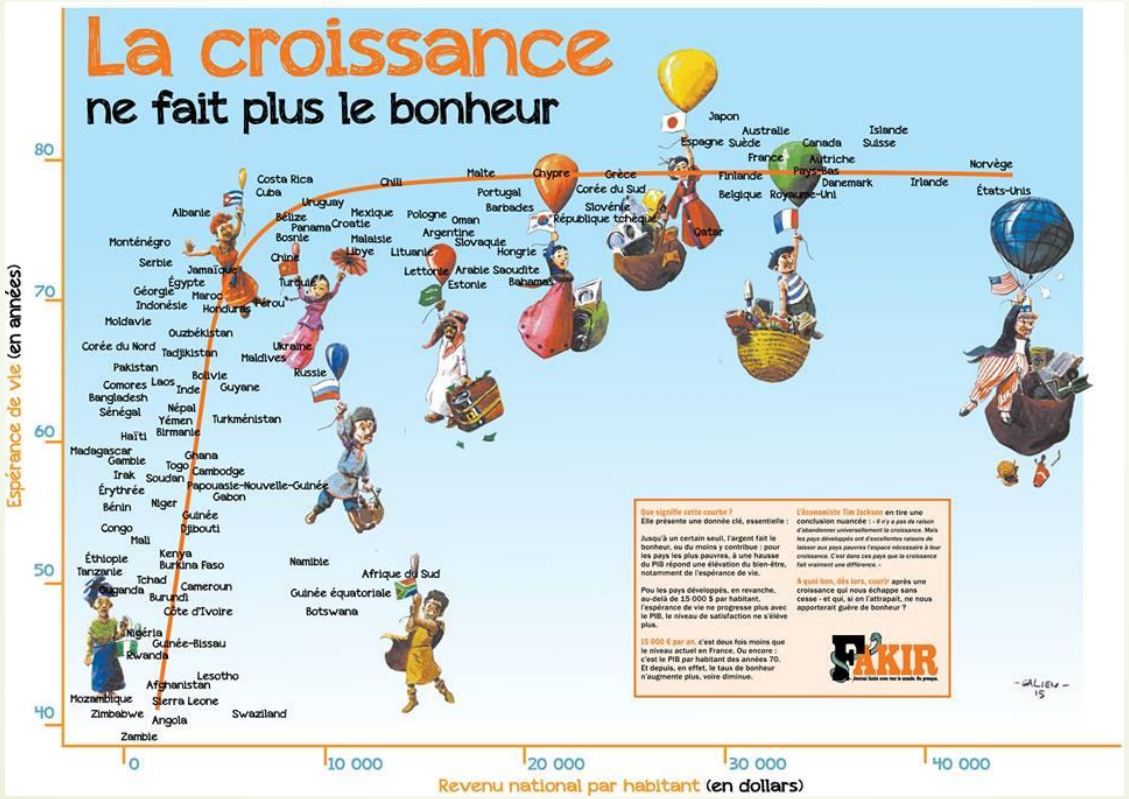
1. Le lien entre deux variables: de la visualisation à l'interprétation causale



1. Le lien entre deux variables: de la visualisation à l'interprétation causale



1. Le lien entre deux variables: de la visualisation à l'interprétation causale



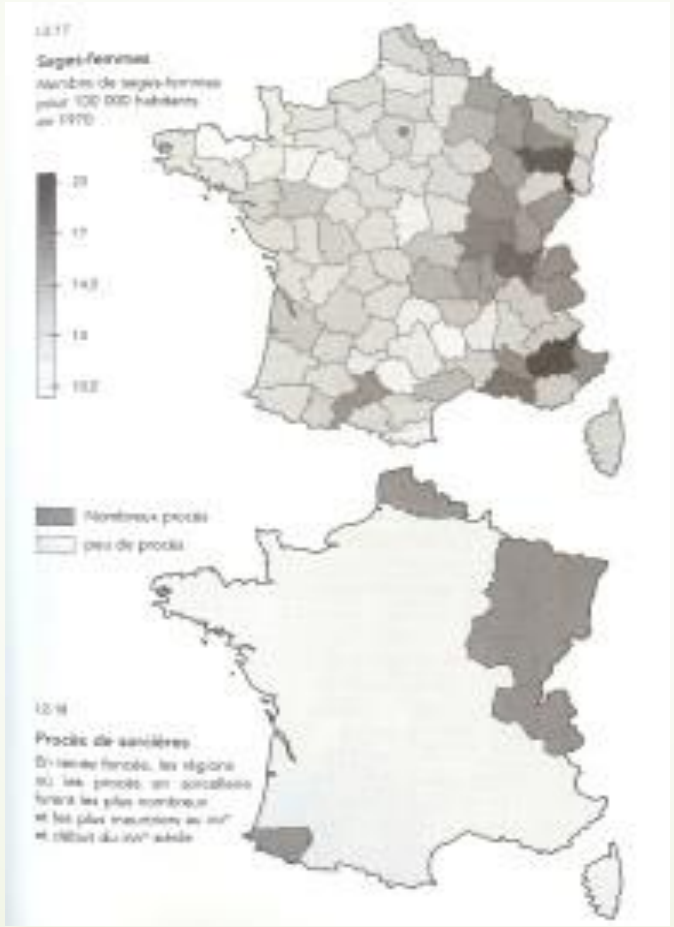
1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- Toutes les relations ne sont pas linéaires. Dans le cas des relations entre bonheur et richesse, on observe souvent des relations logarithmiques. On peut imaginer toutes sortes de relations, la difficulté étant… de les interpréter.
- La construction d'indicateurs permet de construire ce type de mise en relation: « la démocratie favorise la croissance », etc.

1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- La lecture et l'interprétation des diagrammes de dispersion suppose de bien connaître les données et n'est pas toujours simple.
- A nouveau, on suggère des causalités, mais on montre des corrélations, plus ou moins fortes.
- Le problème des « spurious correlations » (« fausses corrélations »),*ie* des corrélations qui ne signifient rien de causal ou même de « lié » par un lien autre que contingent. Exemple des sages-femmes au XXème et des sorcières au XVIème-XVIIème

1. Le lien entre deux variables: de la visualisation à l'interprétation causale



1. Le lien entre deux variables: de la visualisation à l'interprétation causale

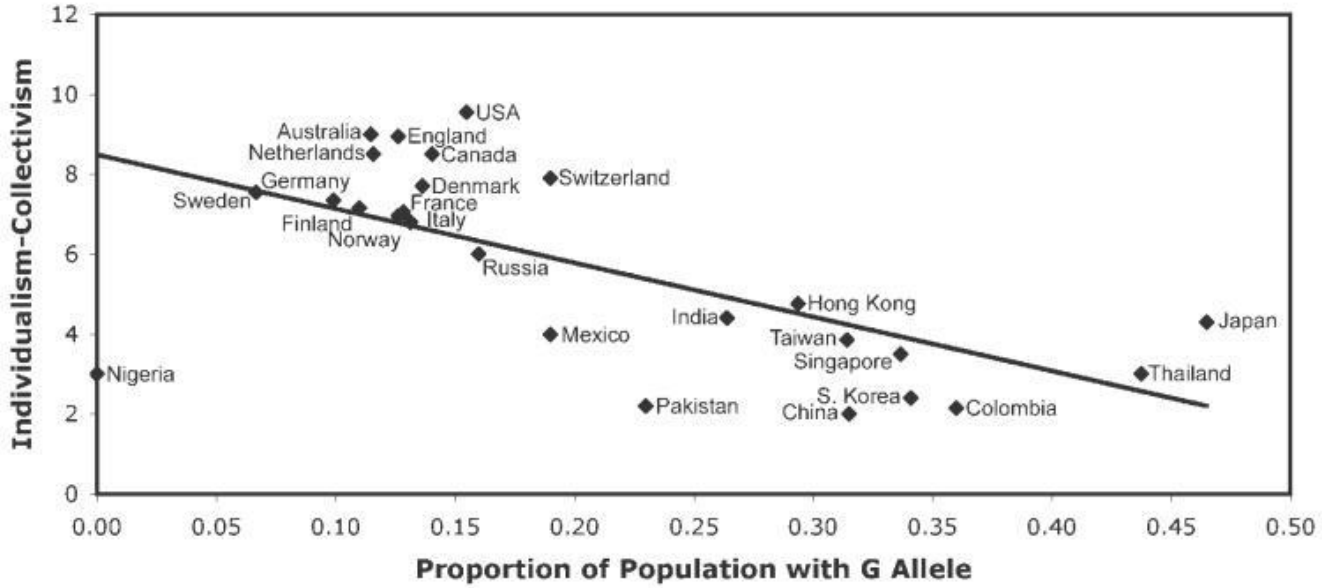
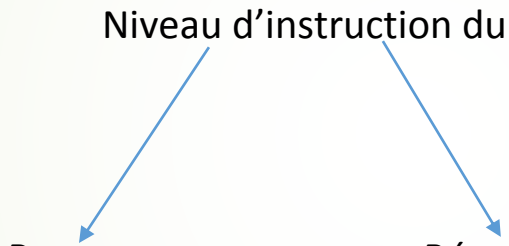


Fig. 1 Correlation between the proportion of the population with the G allele of the A118G polymorphism and individualism-collectivism [Suh *et al.*, 1998; $r(26) = 0.65$, $P < 0.001$]; higher scores represent greater individualism and lower collectivism.

1. Le lien entre deux variables: de la visualisation à l'interprétation causale

- Une pratique omniprésente dans l'espace public: la mise en relation de deux phénomènes sous la forme d'indicateurs. Le débat politique s'appuie en permanence sur ce type de mise en relation. Exemple: Temps de travail ou coût du travail d'une part et emploi de l'autre.
- Nécessité d'un schéma causal, cf. Boudon 1969, à propos des travaux d'Alain Girard sur la réussite sociale. Les résultats scolaires sont corrélés avec les revenus des parents… Mais la variable sous-jacente est le niveau d'instruction des parents.
- On parle ici d'**effet de structure**.

1. Le lien entre deux variables: de la visualisation à l'interprétation causale



Explication de la relation entre revenu de la famille et réussite scolaire de l'enfant (R.Boudon, 1969, p. 62).

2. Le coefficient de corrélation bivariée et la régression linéaire simple

- On cherche un indice du sens et de l'intensité de la relation entre deux variables numériques (x^I, y^I). A l'origine, on voulait mettre en relation l'intelligence et des facteurs héréditaires (Galton et Pearson).
- La **covariance**: moyenne du produit des écarts à la moyenne.

$$Cov\ (x^I, y^I) = \sum f(x^i\ - \bar{x})(y^i - \bar{y})$$

- Pour normaliser la covariance (de sorte qu'elle prenne des valeurs comprises entre -1 et +1), on la divise par le produit des deux écarts-types (des variables). On obtient ainsi le **coefficient de corrélation bivariée** (dit coefficient de Bravais-Pearson), noté r.

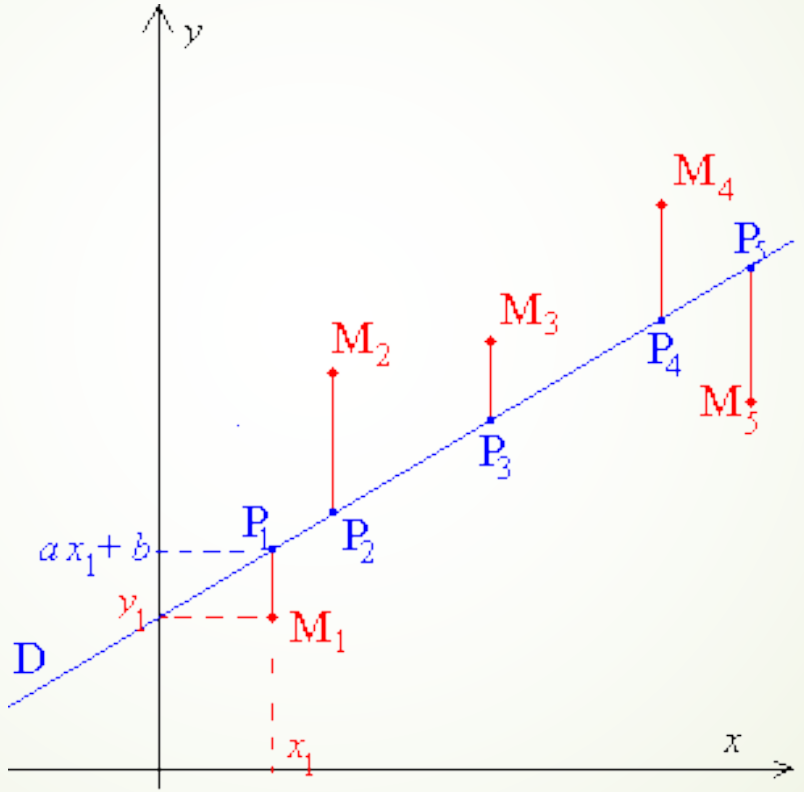
$$r = \frac{Cov\ (x, y)}{Ety\ x\ Ety\ y}$$

- Interprétation: 0 corrélation nulle, $+1$ forte corrélation positive, -1 forte corrélation négative. La valeur est comprise entre -1 et $+1$

2. Le coefficient de corrélation bivariée et la régression linéaire simple

- Régression linéaire simple: on cherche à **ajuster** le nuage de point par une droite qui le « représente » le mieux. Critère : **moindres carrés ordinaires**. On minimise la somme des carrés des distances entre le point d'origine et le point projeté.
- Autre présentation: on cherche la droite telle que, par projection parallèle à l'axe des ordonnées, le nuage projeté ait la variance maximale.
- Une solution, qui passe par le point moyen, telle que $y = ax + b$, avec $a = \frac{Cov\,(x,y)}{Var\,x}$
- On retrouve le coefficient de corrélation r, dont le carré est le coefficient de détermination (r^2), compris entre 0 et 1. Il mesure la part de la variance expliquée par la variable x. Plus il est proche de 1, plus on est proche d'une relation linéaire.
- Variable dépendante, variable indépendante.

2. Le coefficient de corrélation bivariée et la régression linéaire simple



3. La liaison entre deux variables catégorisées : les tableaux de contingence

- On s'intéresse maintenant aux relations entre deux variables qualitatives: c'est le domaine des « tableaux croisés » ou « tableaux de contingence ». Ce tableau nous donne la distribution des effectifs ou des fréquences dans l'ensemble-produit (croisement) de deux variables qualitatives.
- Exemples classiques : relation entre la profession du père et la profession du fils (table de mobilité) ; profession du mari et profession du père de la femme (table d'homogamie).
- Pour Lazarsfeld, le tableau de base de la sociologie.
- On va s'intéresser en particulier à deux éléments:

- La **liaison locale** entre deux modalités (qui équivaut aux notions de sur et sous-représentation)
- La **liaison globale** entre les deux variables (on cherche un indice comparable au coefficient de corrélation bivarié pour les variables qualitatives).

3. La liaison entre deux variables catégorisées : les tableaux de contingence

- La liaison locale entre deux modalités: elle est peu visible à partir du tableau des effectifs observés, donc on va construire des tableaux de fréquences conditionnelles (ou tableaux de pourcentage, en lignes ou en colonnes).
- Une fois construit le tableau de pourcentage, on compare le pourcentage dans la case qui nous intéresse (couple de modalités d'intérêt) au pourcentage correspondant pour l'ensemble de la population (fréquence marginale).
- Si la fréquence dans la case est supérieure, on parlera de liaison positive ou attraction, si elle est égale de liaison nulle, si elle est négative de liaison négative ou répulsion.
- Une mesure de l'intensité de la liaison est donnée par le taux de liaison: $t^{jk} = \frac{f_k^j - f_k}{f_k} = \frac{f_j^k - f_j}{f_j}$ (on soustrait la fréquence de la case par la fréquence marginale et on divise par la fréquence marginale).

3. La liaison entre deux variables catégorisées : les tableaux de contingence

Données Brutes de la mobilité masculine (en milliers)							
	CSP du Fils en 2003						
CSP du père	Agriculteur	Artisan	Cadre	PI	Employé	Ouvrier	**Ensemble**
Agriculteur	252	72	105	190	98	426	**1143**
Artisan	6	182	189	205	79	210	**871**
Cadre	2	37	310	152	37	52	**590**
PI	2	60	266	263	73	135	**799**
Employé	3	43	144	179	108	169	**646**
ouvrier	20	225	304	701	375	1373	**2998**
Ensemble	**285**	**619**	**1318**	**1690**	**770**	**2365**	**7047**

3. La liaison entre deux variables catégorisées : les tableaux de contingence

Fréquence observée

Table de destinée des hommes							
	CSP du Fils en 2003						
CSP du père	Agriculteur	Artisan	Cadre	PI	Employé	Ouvrier	**Ensemble**
Agriculteur	22%	6%	9%	17%	9%	37%	100%
Artisan	1%	21%	22%	24%	9%	24%	100%
Cadre	0%	6%	53%	26%	6%	9%	100%
PI	0%	8%	33%	33%	9%	17%	100%
Employé	0%	7%	22%	28%	17%	26%	100%
ouvrier	1%	8%	10%	23%	13%	46%	100%
Ensemble	4%	9%	19%	24%	11%	34%	100%

Fréquence marginale

Taux de liaison (0,22-0,04)/0,04**=+4,5**

3. La liaison entre deux variables catégorisées : les tableaux de contingence

- On peut construire un graphe des liaisons, un tableau des taux de liaison, etc.
- On veut maintenant mesurer la **liaison globale** entre les deux variables.
- Le raisonnement est simple: on construit un **tableau d'effectifs théoriques** correspondant à la situation d'**indépendance** entre les deux variables. Dans ce cas, on retrouve dans chaque ligne les pourcentages de la ligne marginale (resp. pour le tableau des pourcentages en colonne).

3. La liaison entre deux variables catégorisées : les tableaux de contingence

	CSP fils						
CSP Pères	Ag	Art	Cad	PI	Emp	Ouv	Ens
Ag	46,2260536	100,399745	213,775223	274,112388	124,891443	383,595147	1143
Art	35,2256279	76,5075919	162,903079	208,881794	95,1709947	292,310912	871
Cad	23,8612175	51,82489	110,347666	141,492834	64,4671491	198,006244	590
PI	32,3137505	70,1831985	149,436924	191,614872	87,3038172	268,147439	799
Emp	26,1260111	56,7438626	120,821342	154,922662	70,586065	216,800057	646
Ouv	121,247339	263,340712	560,715766	718,975451	327,580531	1006,1402	2998
Ens	285	619	1318	1690	770	2365	7047

3. La liaison entre deux variables catégorisées : les tableaux de contingence

- On compare ensuite les deux tableaux d'effectifs: on calcule les écarts absolus à l'indépendance case par case $(obs - théo)$.
- Pour obtenir un indicateur de liaison, on élève chaque écart au carré et on le divise par l'effectif théorique. C'est la **distance du $\mathbf{X^2}$**.
- $\mathbf{X^2} = \sum \frac{(obs - théo)^2}{théo}$
- On fait la somme de l'ensemble des distances du $\mathbf{X^2}$ locales pour obtenir la distance du $\mathbf{X^2}$ du tableau. Elle dépend manifestement de la taille de la population (n).
- Pour obtenir un indice descriptif qui ne dépend pas de la taille de la population, on divise $\mathbf{X^2}$ par n.
- $\Phi^2 = \frac{\mathbf{X^2}}{n} = \sum \sum f_j f_k (t^{jk})^2$

3. La liaison entre deux variables catégorisées : les tableaux de contingence

	CSP fils						
CSP Pères	Ag	Art	Cad	PI	Emp	Ouv	Ens
Ag	**915,996795**	8,03334207	55,348085	25,8101938	5,79022627	4,68768071	1015,66632
Art	24,2476111	**145,45809**	4,18070224	0,07213803	2,74769715	23,1776715	199,88391
Cad	20,0288536	4,24076856	**361,231516**	0,78025535	11,702771	107,662379	505,646544
PI	28,4375368	1,47752645	90,9209751	**26,59416**	2,34353083	66,1137787	215,887508
Emp	20,4704953	3,32888442	4,44664955	3,74198449	**19,8311456**	10,5389522	62,3581116
Ouv	84,5463808	5,58216089	117,533675	0,44941287	6,86428483	**133,764769**	348,740683
Ens	1093,72767	168,120772	633,661603	57,4481446	49,2796557	345,945231	2348,18308

3. La liaison entre deux variables catégorisées : les tableaux de contingence

- Dans le cas précis, la distance globale du $\mathbf{X^2}$ est de 2348,18 et l'indice
- $\Phi^2 = \frac{2348,18}{7047} = 0,33$ (plus il s'éloigne de 0 plus la liaison est forte)
- On ne peut comparer que des Φ^2 de tableaux de même format (en toute rigueur, de mêmes marges).

3. La liaison entre deux variables catégorisées : les tableaux de contingence

- Une illustration dans l'histoire de la statistique. Les données « Yeux et Cheveux » (source: Rouanet, Le Roux, 1993).
- Une relation fortement unidimensionnelle entre les deux variables (biologiques).

Le *tableau de contingence* ci–après concerne 5387 enfants écossais du Comté de Caithness répartis selon deux variables catégorisées[3] : la couleur des *Cheveux* (I) à 5 modalités et la couleur des *Yeux* (J) à 4 modalités (données de Tocher).

On se propose d'étudier la *structure de la liaison* entre la couleur des *Yeux* et la couleur des *Cheveux*.

			Yeux				
			Bleu	*Clair*	*Marron*	*Noir*	
			$j1$	$j2$	$j3$	$j4$	
	(Blond)	$i1$	326	688	343	98	1455
	(Roux)	$i2$	38	116	84	48	286
Cheveux	(Châtain clair)	$i3$	241	584	909	403	2137
	(Châtain foncé)	$i4$	110	188	412	681	1391
	(Brun)	$i5$	3	4	26	85	118
			718	1580	1774	1315	5387

Tableau 7.2. Effectifs conjoints et effectifs marginaux.

Les tableaux suivants sont les tableaux des fréquences conditionnelles ou tableaux de transition. Par exemple, la fréquence de $j3$ si $i1$ est $f_{j3}^{i1} = 343/1455 = 0.236$.

	$j1$	$j2$	$j3$	$j4$	
$i1$	.224	.473	.236	.067	1
$i2$	.133	.406	.294	.168	1
$i3$	.113	.273	.425	.189	1
$i4$	.079	.135	.296	.490	1
$i5$	.025	.034	.220	.720	1
	.133	.293	.329	.244	

	$j1$	$j2$	$j3$	$j4$	
$i1$	.454	.435	.193	.075	.270
$i2$	.053	.073	.047	.037	.053
$i3$	.336	.370	.512	.306	.397
$i4$	.153	.119	.232	.518	.258
$i5$	.004	.003	.015	.065	.022
	1	1	1	1	

Tableau 7.5. Tableaux des transitions de I vers J et de J vers I.

Sur le tableau de transition de I vers J, lu comme un tableau de colonnes (variables), on remarque que, pour les yeux $j1$ (*Bleu*) et $j2$ (*Clair*), les fréquences conditionnelles vont en décroissant de $i1$ (Blond) à $i5$ (Brun), alors que pour $j4$ (*Noir*), elles vont en croissant.

Sur le tableau de transition de J vers I, on remarque que, pour $i1$, les fréquences conditionnelles vont en décroissant de $j1$ à $j4$, alors que c'est (à peu près) l'inverse pour $i5$. On remarque aussi la similitude des profils (colonnes) de $j1$ et de $j2$.

Le *carré moyen de contingence* Φ^2 vaut 0.230191.

Les tableau des *taux de liaison* et le *graphe des attractions* (on joint par un trait les modalités en attraction) sont donnés ci–après.

	*j*1	*j*2	*j*3	*j*4
*i*1	0.681	0.612	−0.284	−0.724
*i*2	−0.003	0.383	−0.108	−0.312
*i*3	−0.154	−0.068	0.292	−0.227
*i*4	−0.407	−0.539	−0.101	1.006
*i*5	−0.809	−0.884	−0.331	1.951
	Bleu	*Clair*	*Marron*	*Noir*

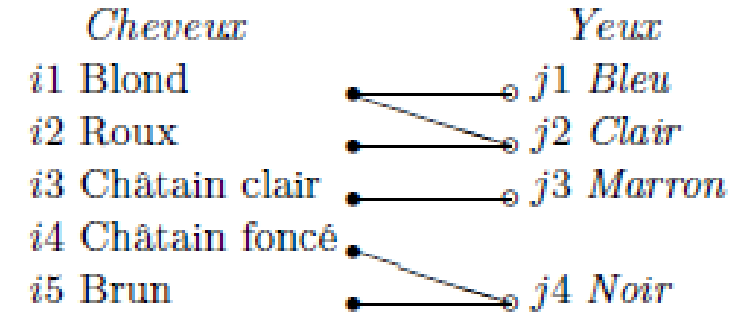


Tableau 7.4. Taux de liaison et graphe des attractions

4. De la description à l'inférence statistique

- Nous revenons maintenant au problème de l'inférence statistique, abordé lors du chapitre consacré aux sources (chapitre 2) et dans les séances précédentes.
- Dans de nombreux cas, on ne travaille que sur de petits **échantillons** des données (échantillon aléatoire, méthode des quotas…).
- On voudrait avoir au moins une idée du **potentiel inductif des données**, et, dans de nombreux cas, on voudrait s'assurer que les effets observés (liaison globale non nulle) ne sont pas dus au hasard d'un petit échantillon très particulier.

4. De la description à l'inférence statistique

- **Cas général**: on se demande si les deux variables sont indépendantes, plus précisément si l'on peut avec une grande certitude écarter l'hypothèse que l'effet observé est dû au hasard. Exemple: le coefficient de corrélation est-il ***significativement*** différent de 0 ?
- Nous examinerons le cas de la relation entre deux variables qualitatives. On a vu que la distance globale du $\mathbf{X^2}$ donnait une mesure de l'écart du tableau à l'indépendance. Cette mesure dépend de la taille de l'échantillon. C'est une statistique inférentielle. On en fait notre statistique de test.

4. De la description à l'inférence statistique

- Plus le $\mathbf{X^2}$ est élevé, plus on s'écarte de l'indépendance. Quand peut-on dire qu'on s'en écarte *significativement* ?
- A partir du moment où le $\mathbf{X^2}$ dépasse une valeur-critique obtenue dans la table de la distribution du $\mathbf{X^2}$ (loi du $\mathbf{X^2}$).
- Pour trouver cette valeur, on utilise deux paramètres: le seuil de significativité a (que l'on peut , par exemple, fixer à $\mathbf{0{,}05}$ ou $\mathbf{0{,}01}$) et le nombre de degrés de liberté $((J-1)X(K-1))$.
- On compare les deux valeurs (observée et valeur-critique) et l'on conclut: l'effet est significatif au seuil de 1%.
- La grandeur du seuil tel que la valeur observée et la valeur-critique sont égaux est appelée $p-value$, notée p.

4. De la description à l'inférence statistique

ν / α	0.1	0.05	0.025	0.01	ν / α	0.1	0.05	0.025	0.01
1	2.71	3.84	5.02	6.63	**16**	23.54	26.30	28.84	32.00
2	4.61	5.99	7.38	9.21	**17**	24.77	27.59	30.19	33.41
3	6.25	7.81	9.35	11.34	**18**	25.99	28.87	31.53	34.80
4	7.78	9.49	11.14	13.28	**19**	27.20	30.14	32.85	36.19
5	9.24	11.07	12.83	15.09	**20**	28.41	31.41	34.17	37.57
6	10.64	12.59	14.45	16.81	**21**	29.61	32.67	35.48	38.93
7	12.02	14.07	16.01	18.47	**22**	30.81	33.92	36.78	40.29
8	13.36	15.51	17.53	20.09	**23**	32.01	35.17	38.08	41.64
9	14.68	16.92	19.02	21.67	**24**	33.20	36.41	39.37	42.98
10	15.99	18.31	20.48	23.21	**25**	**34.38**	**37.65**	**40.65**	**44.31**
11	17.27	19.67	21.92	24.72	**26**	35.56	38.88	41.92	45.64
12	18.55	21.03	23.34	26.22	**27**	36.74	40.11	43.19	46.96
13	19.81	22.36	24.74	27.69	**28**	37.92	41.34	44.46	48.28
14	21.06	23.68	26.12	29.14	**29**	39.09	42.56	45.72	49.59
15	22.31	25.00	27.49	30.58	**30**	40.26	43.77	46.98	50.89

4. De la description à l'inférence statistique

- Dans le cas de notre tableau de mobilité, on a $\nu = 5\text{x}5 = 25$ degrés de liberté et la valeur observée est de $2348{,}18$ (en fait en milliers). Le test est donc significatif au seuil $0{,}01$ (où la valeur-critique est de $44{,}31$).
- On utilise souvent des * pour indiquer le degré de significativité, $*: 0.05, **: 0.01, ***: 0.001$.
- On écrit $p < 0.001$. Parfois, on donne directement la valeur de p telle que la valeur-critique est égale à la valeur observée.

4. De la description à l'inférence statistique

- A quoi cela sert-il ?
- Pas à établir une **causalité**. Celle-ci dépend du schéma explicatif que l'on se donne. Si l'on fait l'hypothèse d'une relation causale, une relation significative est, cependant, couramment utilisée comme **argument** en faveur de cette hypothèse.
- Rôle dans les publications: un garde-fou dans les cas où les échantillons sont faibles ou très faibles.
- Un effet non significatif n'est pas la preuve d'une absence d'effet. Il peut être dû au faible nombre d'observations.
- Dire que l'effet est significatif ne dit rien sur l'**importance** de l'effet. La significativité dépend de l'importance de l'effet *et* de la taille de l'échantillon. Il faut d'abord s'assurer descriptivement que l'effet est suffisamment important ou « notable ».

Conclusion

- Trois niveaux distincts d'analyse:

⇒ liaison ou corrélation dans l'échantillon étudié (description) ;

⇒ significativité (inférence) ;

⇒ causalité (schéma causal).

- En sciences sociales, peu de phénomènes se réduisent au jeu de deux variables (une explicative une expliquée): les processus sont multidimensionnels, d'où la nécessité de méthodes multivariées et multidimensionnelles.
- La structure des relations entre les variables est complexe et nécessite des interprétations subtiles.