nent les résultats des deux instituts ne sont pas contradictoires, mais aucun ;ux ne permet de conclure à l'avantage d'un des deux candidats au moment ıterrogation. D'où l'importance centrale de l'utilisation de l'intervalle de nce, en particulier pour des proportions proches de 50% !

s articles d'application conseillés

Pascal Ardilly, « Nature et déterminants de l'erreur d'échantillonnage dans les enquêtes par sondage », *Statistique et société*, vol. 1, n°2, 2013, p. 43-50.
Jacques Poitevineau, « L'usage des tests statistiques par les chercheurs en psychologie : aspects normatif, descriptif et prescriptif », *Mathématiques et sciences humaines*, vol. 167, n°3, 2004, p. 5-25.

Chapitre 5

La mesure du lien statistique entre deux variables quantitatives

Par ce chapitre, nous entrons de plain-pied dans l'analyse bidimensionnelle. Jusqu'à présent et outre les premiers développements portant sur la démarche générale de la récolte de l'information à traiter, nous nous sommes davantage focalisés sur la description des variables appréhendées séparément (si ce n'est la question de la construction d'indicateurs). L'objectif était de comprendre l'importance de connaître parfaitement les détails de son fichier de données, de détecter d'éventuelles erreurs de codage, et d'apporter quelques premiers éléments de description. Rappelons que ces étapes sont indispensables pour maîtriser et préparer son fichier de données en vue d'analyses statistiques ultérieures plus avancées.

Il s'agit désormais, dans les chapitres à venir, de se pencher sur les liens supposés, par hypothèses de travail, exister entre deux variables. Nous abordons ici la question délicate de la causalité, sur laquelle nous reviendrons dans ce chapitre. Lorsque nous nous posons une question de sciences sociales par une démarche quantitative bivariée, l'hypothèse sous-jacente consiste le plus souvent à supposer qu'il existe un lien (de causalité) entre les deux variables que l'on confronte. Qu'entendons-nous par-là ? Qu'il existe des logiques sociales entre ces deux variables, qu'elles évoluent ensemble de façon logique, simultanée, ou que le fait de se trouver dans une situation donnée favorise la présence concomitante d'une seconde situation. On se propose alors dans un premier temps d'en apporter la preuve statistique, pour ensuite apporter des éléments explicatifs substantiels, d'ordre sociologique. Nous insisterons ici sur la première phase, qui consiste à démontrer l'existence d'un lien statistique ; au sociologue, au démographe, à l'économiste, etc. d'apporter ensuite la preuve d'une cohérence théorique, conceptuelle et intellectuelle.

Ici, la démarche méthodologique dépend du type de variables en présence. Le tableau 1 en résume les grands principes. Ce chapitre traitera de la mise en relation de deux variables quantitatives (les suivants des autres cas de figure).

Le lien statistique entre deux variables quantitatives est mesurable à l'aide du coefficient de corrélation linéaire et consiste à confirmer statistiquement l'existence d'une variation simultanée entre les deux. Nous exposerons tout d'abord la représentation graphique adaptée à l'étude du lien entre deux variables de ce type, puis expliquerons comment la décrire et l'interpréter à l'aide du coefficient de corrélation linéaire. Enfin, nous présenterons les limites inhérentes à ce coefficient et comment y remédier lorsque cela est possible.

Tableau 1 : Quel type d'analyse bivariée pour quel type de variable ?

		Variable à expliquer				
		Nominale	Ordinale		Quantitative	
	Nominale	Tableau croisé Test du Chi-deux → Chapitre 6	Tableau croisé Test du Chi-deux → Chapitre 6		Test de Student Analyse de variance → Chapitre 7	
ıiable ıicative	Ordinale	Tableau croisé Test du Chi-deux → Chapitre 6	Tableau croisé Test du Chi-deux → Chapitre 6	*Nuage de points** *Analyse de corrélation** *→ Chapitre 5**	Test de Student Analyse de variance → Chapitre 7	*Nuage de points** *Analyse de corrélation** *→ Chapitre 5**
	Quantitative	Après recodage de la variable quantitative Tableau croisé / Test du Chi-deux → Chapitre 6	Après recodage de la variable quantitative Tableau croisé / Test du Chi-deux → Chapitre 6	*Nuage de points** *Analyse de corrélation** *→ Chapitre 5**	Nuage de points Analyse de corrélation → Chapitre 5	

limites possibles si la ou les variable(s) ordinale(s) se présentent sous forme d'échelles (de 0 à 10 ou de 1 à 10 par exemple)

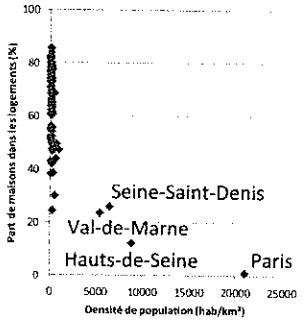
nuage de points

)mme détaillé dans le chapitre 8, le nuage de points est particulièrement té pour visualiser graphiquement l'éventuel lien entre deux variables titatives et, le cas échéant, mesurer l'intensité de ce lien. Nous verrons dans te de ce chapitre que la forme du nuage de points et la manière dont ils se tissent dans le graphique sont un premier indicateur de ce lien.

a lecture d'un nuage de points à partir d'exemples

enons un premier exemple représentant le lien entre la densité de lation et la part de l'habitat individuel dans les logements, par département e 1a). Un point du nuage représenté correspond à une unité statistique, soit :s 96 départements de France métropolitaine pour cet exemple[1].

ıre 1a : Densité de population et part de l'habitat individuel dans les logements, par département, France métropolitaine

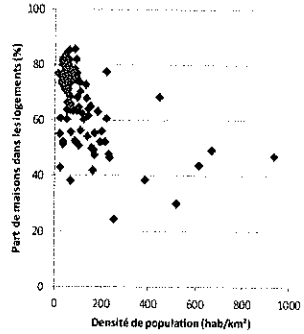


Sources : INSEE, 2008

marque de construction d'un nuage de points : lorsqu'une des deux variables a le rôle de iable explicative et l'autre de variable à expliquer (c'est-à-dire lorsque l'on recherche une ation de causalité), la première doit être placée sur l'axe des abscisses et la seconde sur l'axe s ordonnées. Mais ces rôles ne sont pas toujours clairement définis : c'est le cas par exemple sque les deux variables tiennent des rôles symétriques.

En première lecture, nous constatons qu'il existe quatre valeurs aberrantes (ou *outliers* : les points correspondant à une densité supérieure à 5000 hab/km^2) qui écrasent le nuage de points et ne permettent pas de visualiser la forme de celui-ci. Dans cette situation, il s'avère utile d'identifier à quels départements ils correspondent et de représenter à nouveau le nuage de points, en les excluant (figure 1b). Il s'agit, dans l'ordre décroissant des densités de population, de Paris (21060 hab/km^2), des Hauts-de-Seine (8805 hab/km^2), de la Seine-Saint-Denis (6383 hab/km^2) et du Val-de-Marne (5351 hab/km^2), soit Paris et les trois départements de la petite couronne parisienne (très spécifiques quant à leur structure d'habitat).

Figure 1b : Densité de population et part de l'habitat individuel dans les logements par département, France métropolitaine, hors Paris et petite couronne



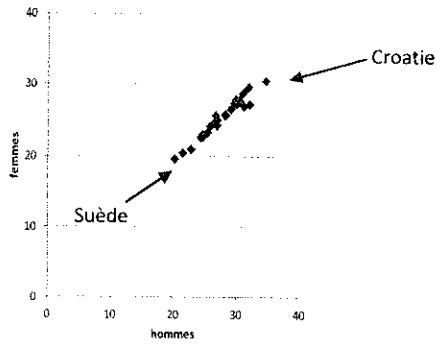
Sources : INSEE, 2008

Malgré la suppression des valeurs extrêmes, ce nouveau nuage de points reste difficile à interpréter à lui tout seul du point de vue du lien entre les deux variables étudiées. On remarque en effet que globalement, plus la densité de population est élevée, plus la part de maisons dans les logements est faible, mais il est impossible d'aller plus loin dans l'interprétation, en particulier sur l'intensité de ce lien.

Prenons un second exemple : la représentation du lien entre l'âge moyen estimé des jeunes quand ils quittent le domicile parental pour les hommes et pour les femmes dans 28 États membres de l'Union Européenne (UE)[1]. Ces deux variables sont totalement symétriques et l'on peut donc placer indifféremment l'une ou l'autre sur l'axe des abscisses et inversement (figure 2). Ici un point correspond à un des 28 pays États membres de l'UE.

1 Les données correspondant à cet exemple sont présentées en détail dans le chapitre 2.

ıre 2 : Âge moyen estimé des jeunes quand ils quittent le domicile parental pour les hommes et pour les femmes, 28 États membres de l'Union Européenne



Sources : Eurostat, 2012

s points de ce nuage se répartissent visuellement sur une quasi ligne droite. :marque ainsi que si l'âge moyen estimé des jeunes quand ils quittent le :ile parental est élevé pour l'un des deux sexes, il le sera également pour e, et inversement. Ici, il semble donc possible de dégager des logiques ıales spécifiques.

ı forme et la dispersion d'un nuage de points

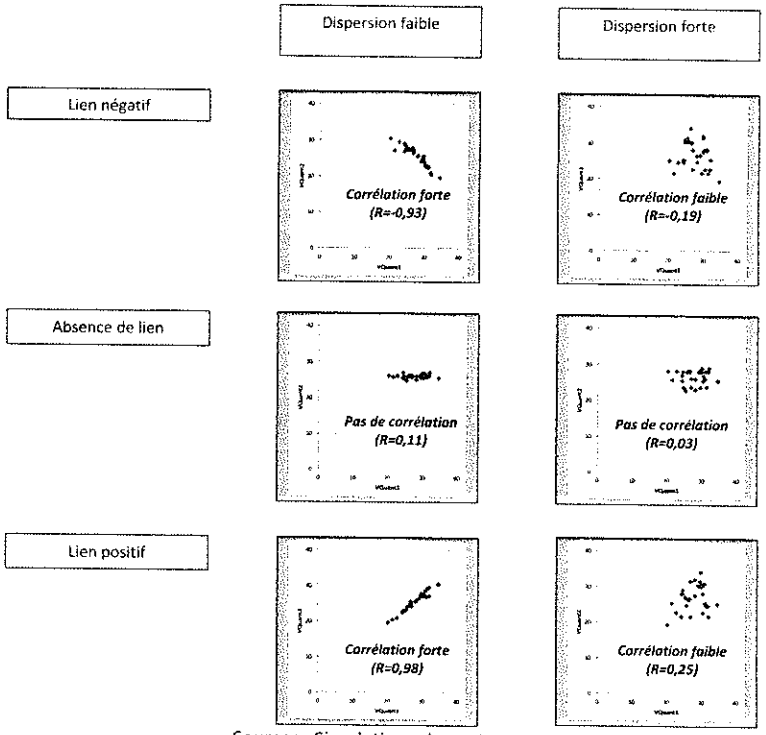
partir des représentations 1b et 2, quelques règles de base pour la lecture ıuage de points se dégagent d'ores et déjà :

la forme du nuage nous indique à la fois la présence, effective ou non, d'un potentiel lien entre les deux variables et la direction de celui-ci (décroissant dans la figure 1b, croissant dans la figure 2) ;

la plus ou moins grande dispersion (ou éparpillement) des points nous indique l'intensité de ce lien (faible dans la figure 1b, fort dans la figure 2).

figure 3 résume les différentes possibilités de lecture d'un nuage de points.

Figure 3 : Lien en fonction de la forme et de la dispersion du nuage de points



Sources : Simulations des auteurs

II. La mesure statistique : le coefficient de corrélation linéaire

Si le nuage de points permet de visualiser graphiquement un éventuel lien entre deux variables, il ne permet cependant pas de le quantifier, c'est-à-dire d'en évaluer précisément la force. Il est donc nécessaire de calculer un coefficient chiffrant l'intensité de ce lien, afin d'établir statistiquement sa puissance ou sa relative faiblesse.

II.1. La formule : le R de Pearson

Le ***coefficient de corrélation linéaire de Pearson*** (généralement noté R) entre deux variables quantitatives X_1 et X_2 s'emploie à déterminer l'intensité et la direction du lien existant entre elles. Il se calcule de la manière suivante :

$$R = \frac{\sum_{i=1}^{n}(x_{1i} - \overline{X_1})(x_{2i} - \overline{X_2})}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \overline{X_1})^2} \times \sqrt{\sum_{i=1}^{n}(x_{2i} - \overline{X_2})^2}}$$

Détaillons cette formule afin de comprendre ce qu'il permet de mesurer.

umérateur calcule, pour une unité statistique donnée et pour chacune des ariables, l'écart entre la valeur observée pour l'unité et la moyenne de ariable. Ces deux écarts sont ensuite multipliés afin de prendre en compte nultanéité, puis les produits ainsi obtenus sont additionnés sur l'ensemble ités statistiques. **Ce numérateur mesure donc au final la variation te des deux variables : quand X_1 prend des valeurs élevées, X_2 prend-elle itamment des valeurs élevées ou faibles, et inversement ? Avec quelle té ?** En d'autres termes, les deux variables évoluent-elles de concert ?

énominateur, quant à lui, est égal au produit des deux écart-types (des ariables en présence). Il permet de normaliser la variation conjointe des riables mesurée par le numérateur, c'est-à-dire de rendre les variations de e comparables entre elles en les rapportant à la même unité de mesure été de l'écart-type déjà abordée au cours du chapitre 2 : il permet de er à partir d'une unique unité de mesure. C'est important pour l'étude du tre deux variables quantitatives, dans la mesure où ne s'exprimant pas airement au départ dans la même unité[1], elles sont ainsi standardisées).

apport de ces deux valeurs quantifie alors la variation conjointe relative ix variables X_1 et X_2.

s caractéristiques du coefficient de corrélation linéaire

oefficient présente un certain nombre de caractéristiques mathématiques tant une interprétation claire (voir figure 3) :

il est borné entre -1 et +1.

Un coefficient R égal à 0 signifie l'absence totale de lien entre les deux variables. Le nuage de points se caractérise alors par une allure sphérique.

Le signe de R indique le sens du lien, c'est-à-dire sa direction (croissante ou décroissante).

Un lien négatif (R<0) signifie que lorsqu'une des deux variables prend des valeurs élevées, l'autre prend des valeurs faibles et inversement. Le nuage de points a alors une allure décroissante.

Un lien positif (R>0) signifie que lorsqu'une des deux variables prend des valeurs élevées, l'autre également et inversement. Le nuage de points prend alors une allure croissante ;

plus le coefficient s'éloigne de 0, plus le lien est fort.

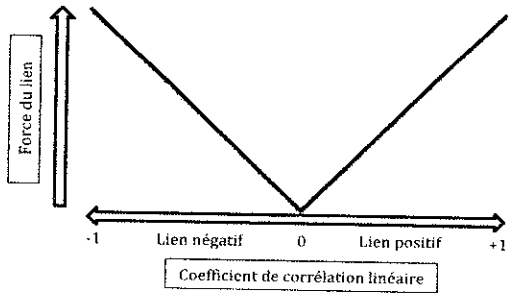
Un coefficient égal à 1 marque un lien positif parfait et un coefficient égal à -1 signifie un lien négatif parfait. Dans ces deux cas de figure « parfaits », les points sont tous alignés sur une droite. Un R proche de 1 en valeur absolue aboutira à un nuage de points peu éparpillé, croissant ou décroissant. Il aura l'allure d'un ballon de rugby de plus en plus écrasé à mesure qu'il se rapproche de 1 en valeur absolue.

La force du lien entre deux variables quantitatives telle que mesurée par le R de Pearson peut se résumer à l'aide de la figure 4.

---

le cas par exemple si l'on étudie les liens entre le revenu en euros (qui peut aller jusqu'à eurs centaines de milliers d'euros) et l'âge en années (qui dépasse rarement les 100 ans). voyons bien que ces deux variables font référence à des ordres de grandeur très différents, ıi est neutralisé par l'écart-type.

Figure 4 : Force du lien entre deux variables quantitatives en fonction du coefficient de corrélation linéaire



II.3. L'interprétation du coefficient de corrélation linéaire à partir d'exemples

Reprenons l'exemple mesurant le lien entre la densité de population et la part de l'habitat individuel dans les logements. Le coefficient de corrélation linéaire correspondant à la figure 1a vaut -0,59 et celui pour la figure 1b -0,52 :

- le coefficient R est négatif, donc le lien est négatif. Cela confirme la première lecture du nuage de points : lorsque la densité de population est forte, la part de maisons dans les logements est faible ;
- il est relativement médian en termes d'intensité. Ce lien négatif existe mais n'est pas parfait. Il serait intéressant de le comparer avec celui obtenu dans d'autres pays par exemple ;
- il diminue légèrement en intensité lorsque l'on retire les quatre *outliers* de Paris et de la petite couronne. Cela confirme le fait que ceux-ci tirent le nuage de points et donc affectent artificiellement l'intensité du lien.

L'exemple de la figure 2 est également très intéressant du point de vue de la corrélation linéaire. En effet ce coefficient R vaut 0,98 :

- R est positif, donc le lien est positif. Cela confirme l'interprétation faite du nuage de points : lorsque dans un pays, l'âge moyen estimé des jeunes au départ du domicile parental est tardif pour les femmes, il l'est également pour les hommes (et s'il est précoce pour l'un, il le sera aussi pour l'autre sexe) ;
- R est très proche de 1. Le lien positif est donc quasi-parfait : les deux variables sont parfaitement corrélées positivement. Attention toutefois (voir également chapitre 2) : cela ne s'interprète pas nécessairement comme une quasi-redondance des deux variables : en effet, cela ne signifie pas que pour un pays donné, hommes et femmes décohabitent aux mêmes âges de chez leurs parents (ce peut être le cas, mais pas nécessairement). Cela indique en revanche que dans un pays relativement aux autres situations nationales, si les hommes quittent le domicile parental à des âges plutôt élevés, alors ce sera également le cas des femmes (il peut donc exister un différentiel genré, qui plus est variable d'un pays à l'autre, des âges de la décohabitation ; c'est d'ailleurs effectivement le cas, comme nous l'avons souligné au chapitre 2). En d'autres termes, si le coefficient de

corrélation informe sur la direction et l'éparpillement des unités statistiques (donc si les points suivent une droite ou non et si cette dernière est croissante ou décroissante), il ne fournit en revanche aucune indication concernant la pente de la droite (plus ou moins inclinée) ou sur sa proximité à la bissectrice[1].

:s limites du coefficient de corrélation linéaire

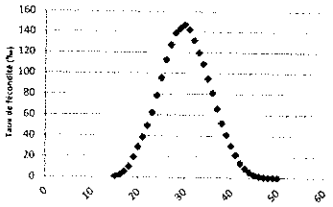
le coefficient de corrélation linéaire, par son bornage et sa symétrie, est nent interprétable, il se confronte néanmoins en pratique à un certain re de limites[2].

a linéarité

ıt d'abord, comme son nom l'indique, il s'agit d'un coefficient adapté à la e d'un lien statistique de type **linéaire**, c'est-à-dire d'un lien statistique nt être résumé à l'aide d'une droite. Tout lien statistique non linéaire sera oire pas du tout mis en évidence par ce coefficient. Il est en effet tout à fait le d'obtenir un coefficient de corrélation linéaire égal à 0 alors même qu'il un lien, mais non linéaire, entre deux variables quantitatives. En cela, ser le nuage de points associé à la corrélation entre deux variables sert de fou : l'observation de sa forme permet directement de percevoir si la n décrite est bien d'ordre linéaire ou non et si le calcul du R de Pearson a du u non.

nons l'exemple du taux de fécondité en fonction de l'âge des mères :ielles en France pour l'année 2013 (figure 5).

ıre 5 : Taux de fécondité en fonction de l'âge des mères potentielles, France hors Mayotte



Sources : INSEE, statistiques de l'état civil et estimations de population

coefficient de corrélation linéaire obtenu s'élève à -0,24. Cette valeur ve tendrait à indiquer un lien statistique linéaire négatif : plus les femmes

ıissectrice est une droite remarquable, sur laquelle les valeurs des deux variables sont tiques. Dans l'exemple, si les âges à la décohabitation s'alignaient sur une bissectrice, cela fierait, que dans chacun des pays, hommes et femmes quittent le domicile parental au ıe âge (cet âge peut néanmoins varier d'un pays à l'autre).

s invitons le lecteur souhaitant se confronter à des exemples complémentaires à ce chapitre reporter à Nicolas Gauvrit, « Chapitre 4 : causalité, corrélation », *Statistiques, méfiez-vous !*, ;, Ellipses, 2007, p. 103-139.

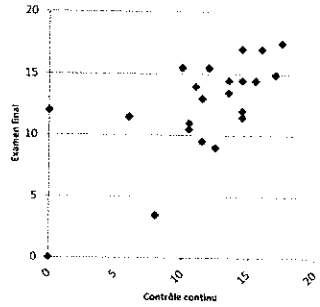
seraient âgées, moins elles auraient d'enfants, ce qui paraît logique au premier abord. Mais attention toutefois : la relative faiblesse du coefficient indique que ce lien est ténu. En réalité, il dissimule la partie gauche du nuage de points : si les femmes plus âgées ont en effet moins d'enfants que les autres, les plus jeunes en ont également moins. Ce nuage de points suit en réalité une courbe en U inversé bien connue maintenant du lecteur : sa forme suit celle d'une loi normale ! Le lien statistique ne peut ici clairement pas être résumé au moyen d'un coefficient de corrélation linéaire puisque le lien observé n'est pas de nature linéaire mais normale, avec une moyenne autour de 30 ans.

III.2. L'analyse des points aberrants (*outliers*)

Nous avons également vu dans l'exemple des figures 1a et 1b de ce chapitre que le coefficient de corrélation peut varier selon la prise en compte ou non des *outliers*. Avec des coefficients s'élevant respectivement à -0,59 et -0,52, le lien devient ici moins fort en supprimant les quatre *outliers* du calcul de corrélation.

Cette sensibilité aux *outliers* peut se révéler encore plus marquée. Ainsi l'exemple d'un groupe d'étudiants et leurs résultats du semestre au contrôle continu et à l'examen final (figure 6)[1].

Figure 6 : Note à l'examen final en fonction de la note au contrôle continu



Sources : Groupe d'étudiants suivi par les auteurs

Le coefficient de corrélation linéaire associé est égal à 0,67. Il est positif et élevé : plus la note obtenue par l'étudiant au contrôle continu est élevée, plus sa note à l'examen l'est aussi. Néanmoins, le lecteur avisé remarque l'existence d'un *outlier* tout particulier : un des points du nuage correspond à un étudiant ayant obtenu 0 au contrôle continu et 0 à l'examen final. Cet étudiant ne s'est jamais présenté à l'enseignement concerné mais était bien inscrit au cours. Le coefficient de corrélation linéaire recalculé sans la prise en compte de cet *outlier* vaut 0,49. Il s'avère bien moins élevé que le précédent et relativise la conclusion apportée plus

[1] L'exemple qui suit est issu d'un cas réel vécu il y a quelques années par les auteurs. Les données sont volontairement anonymisées pour des raisons évidentes de confidentialité.

haut (tout en témoignant d'une relative cohérence entre note au contrôle continu et note à l'examen final[1]).

III.3. L'erreur écologique

Nous avons vu dans l'introduction qu'il n'est pas possible de reconstruire des données individuelles à partir de données agrégées. De cette impossibilité découle un écueil méthodologique que l'on appelle *erreur écologique* : la mesure du coefficient de corrélation dépend de l'échelle à laquelle on choisit de le calculer. Cela signifie qu'un coefficient de corrélation calculé sur des données agrégées ne sera pas toujours du même ordre qu'à l'échelle individuelle. Il pourra même éventuellement aller jusqu'à s'inverser ! Il importe donc de prendre des précautions supplémentaires au moment de l'interprétation d'un coefficient de corrélation lorsque celui-ci est calculé, comme c'est souvent le cas, sur des données agrégées et ne surtout pas formuler à partir de ce dernier des conclusions au niveau individuel.

L'exemple développé par William Robinson dans son article de 1950[2] est une illustration parlante d'erreur écologique et est parfaitement résumé par Nonna Mayer[3] : « Analysant la relation entre race et alphabétisation aux États-Unis, [W. Robinson] observe au niveau des neuf grandes divisions du territoire américain une corrélation presque parfaite entre le taux d'analphabétisation et le poids de la population noire ([R=]0,94). Mais cette corrélation s'atténue à mesure que se réduit la taille de l'unité d'observation, passant de [R=]0,77 au niveau du comté pour tomber à [R=]0,20 au niveau des individus. Les Noirs ne sont pas plus souvent analphabètes que les Blancs, mais ils résident plus souvent dans des zones rurales où le taux d'alphabétisation est plus faible. » Dès lors, attention à ne pas formuler des conclusions individuelles à partir d'un résultat qui se fonde sur des données agrégées !

Les données électorales se prêtent également facilement à l'erreur écologique. Prenons ainsi les résultats du premier tour de l'élection présidentielle de 2012 pour la France métropolitaine. À l'échelle des 22 régions, le coefficient de corrélation entre le score aux exprimés pour Nicolas Sarkozy et celui pour Marine Le Pen vaut R=0,26. Il diminue fortement pour être quasi nul à l'échelle des 96 départements (R=0,05) pour s'inverser à l'échelle des circonscriptions législatives (R=-0,11). L'échelle régionale apparaît clairement peu précise pour mesurer ce type de phénomène. Elle agrège en effet des logiques géographiques et sociodémographiques différentes (degré d'urbanité ou proportion d'ouvriers par exemple).

1 On constate cependant que quelques étudiants ayant obtenu une note supérieure à la moyenne au cours du contrôle continu ont relâché leur effort au moment de l'examen final. Cela contribue à l'obtention d'un coefficient finalement assez peu élevé.

2 William S. Robinson, « Ecological correlations and behavior of individuals », *American sociological review*, n°15, 1950, p. 351-357.

3 Nonna Mayer, *Sociologie des comportements politiques*, Paris, Armand Colin, 2010, p. 72.

Comment pallier à l'erreur écologique ? Il existe plusieurs méthodes d'analyse multivariée que nous ne détaillerons pas ici, comme la modélisation multiniveaux[1] qui permet de prendre en compte plusieurs niveaux d'analyse et ainsi de mesurer leurs effets et leurs possibles interactions.

Encart 1 : ATTENTION PRUDENCE !
Corrélation n'est pas causalité !

Il faut être vigilant lors de l'interprétation du lien constaté entre deux variables, et ce d'ailleurs quels que soient leurs types (cet encart est donc valable pour l'analyse du lien entre deux variables qualitatives, ou entre une variable quantitative et une variable qualitative). En effet, une forte corrélation linéaire ou plus largement un fort lien entre deux variables ne signifient pas nécessairement une relation de cause à effet de l'une vers l'autre. Autrement dit, il existe une nuance importante entre conclusion statistique et conclusion sociologique. Il faut surtout prendre toutes les précautions quand il s'agit de réaliser le passage de la première vers la seconde : si le chiffre statistique est indispensable pour conclure du point de vue interprétatif, seul le sociologue est en mesure d'établir de façon certaine qu'un lien, même avéré statistiquement, ne provient pas d'une coïncidence malencontreuse qui pourrait mener à de grossières erreurs d'interprétation.

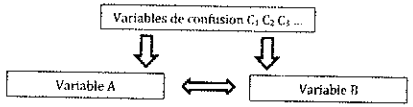
Reprenons l'exemple développé par William Robinson. La corrélation quasi-parfaite entre le taux d'analphabétisation et la proportion de la population noire cache en fait l'intervention d'une troisième variable, le niveau de ruralité (et les conditions socio-économiques) de la zone géographique considérée.

De manière générale, un lien statistique apparent entre une variable A et une autre B peut faire intervenir une variable dite de confusion C, parfois aussi appelée variable latente, corrélée à A d'une part et à B d'autre part. Dans ce cas la corrélation entre A et B cache en réalité une relation plus complexe entre trois variables. Ce raisonnement est bien entendu généralisable au cas où plusieurs

1 « La caractéristique principale de la méthode multiniveaux est de modéliser simultanément l'effet de variables individuelles et contextuelles dans des équations distinctes mais connectées (les paramètres de la première équation devenant les variables dépendantes dans les suivantes [...]). La première équation concerne le niveau micro des individus et les équations suivantes concernent le niveau macro du contexte. Cela permet de prendre en compte l'interaction entre les deux niveaux et, surtout, l'hétérogénéité qui existe entre les contextes. Cela évite ainsi de conclure à tort qu'une variable n'est pas influente parce qu'en fait elle a un impact seulement dans certains lieux ou seulement certaines années. » (Anne Jadot, Marcel Van Egmond, « Réconcilier l'individuel et le contextuel ? L'intérêt de la méthode multiniveaux en recherche électorale », *Revue de la maison française d'Oxford*, vol. 1, n°1, 2003, p. 226).
Pour davantage de précisions sur cette méthode avancée de traitement statistique : Pascal Bressoux, *Modélisation statistique appliquée aux sciences sociales*, Bruxelles, De Boeck, 2008, 464 p.

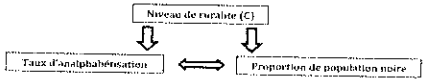
variables de confusion interviennent. La figure 7a schématise ce concept de variables de confusion.

Figure 7a : Corrélation et variables de confusion



L'exemple de William Robinson se schématise ainsi (figure 7b).

Figure 7b : Variable de confusion, exemple de W. Robinson



Il existe également des cas où deux variables peuvent être corrélées sans aucunement être liées entre elles du point de vue de leur sens. Prenons l'exemple du tableau 2 ci-dessous.

Tableau 2 : Consommation de mozzarella et nombre de doctorats délivrés en ingénierie civile (données américaines)[1]

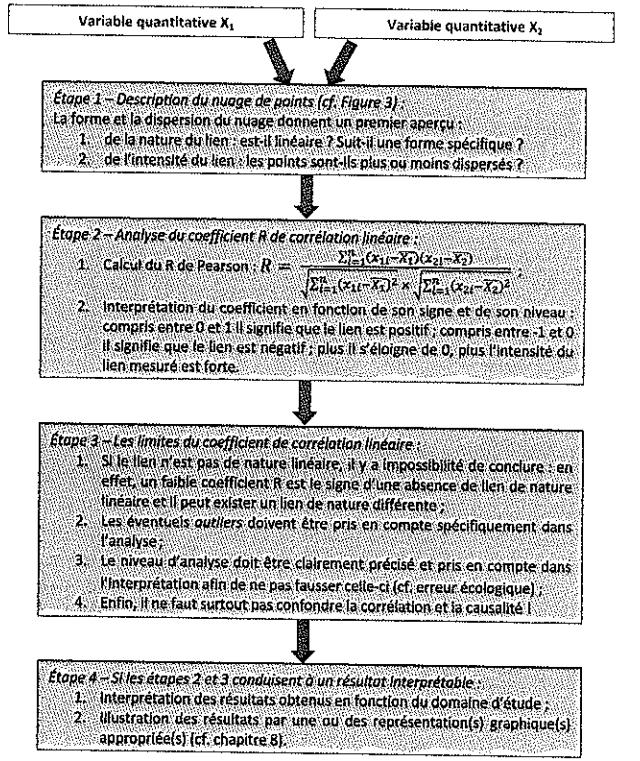
	Per capita consumption of mozzarella cheese, Pounds	Civil engineering doctorates awarded, Degrees awarded
2000	9,3	480
2001	9,7	501
2002	9,7	540
2003	9,7	552
2004	9,9	547
2005	10,2	622
2006	10,5	655
2007	11	701
2008	10,6	712
2009	10,6	708

Sources : USDA et National Science Foundation

Le coefficient de corrélation linéaire entre la consommation de mozzarella et le nombre de doctorats délivrés en ingénierie civile vaut $R = 0{,}96$, soit une corrélation positive presque parfaite. Peut-on pour autant en déduire que le niveau de la consommation en mozzarella influe sur le nombre de doctorats en ingénierie civile délivrés (ou inversement) ? Bien sûr que non ! Ces deux variables évoluent de la même manière mais indépendamment l'une de l'autre, sans relation de causalité.

[1] Le site Internet <http://www.tylervigen.com> (consulté le 13 mai 2014) fournit plusieurs exemples de fortes corrélations ne révélant pas pour autant la présence d'un lien de causalité entre les deux variables prises en compte.

Encart 2 : Schéma récapitulatif des notions abordées dans ce chapitre



IV. Les exercices d'application

IV.1. Le taux de chômage chez les femmes et chez les hommes

Le tableau 3 présente les taux de chômage par sexe dans différents pays d'UE et hors UE.

Tableau 3 : Taux de chômage par sexe et par pays, décembre 2013

Pays	Taux de chômage des hommes	Taux de chômage des femmes
Belgique	8,7	8,1
Bulgarie	14,2	12,1
République tchèque	5,7	8,1
Danemark	6,6	7,6
Allemagne	5,5	4,8
Estonie	9,2	8,2
Irlande	13,4	10,3
Grèce	24,5	31,6
Espagne	25,1	26,7
France	10,2	10,3
Croatie	17,4	17,3
Italie	12,1	13,6
Chypre	18	15,7
Lettonie	12,2	11
Lituanie	12,2	9,7
Luxembourg	5,7	6,7
Hongrie	8,4	9,2
Malte	6,8	6,8
Pays-Bas	7,4	6,6
Autriche	4,8	5,3
Pologne	9,3	10,6
Portugal	15	15,8
Roumanie	7,9	6,5
Slovénie	9,2	10,5
Slovaquie	14,3	13,8
Finlande	9	7,7
Suède	8,1	7,9
Royaume-Uni	7,4	6,8
Islande	5,8	5,1
Norvège	3,8	3,5
Turquie	7,8	10,2
États-Unis	6,8	6,5
Japon	3,8	3,5

Sources : Eurostat, décembre 2013

Représenter graphiquement le nuage de points associé à ce tableau. À quoi correspond un point du nuage ? Calculer ensuite le coefficient de corrélation associé et l'interpréter.

IV.2. La densité de médecins et de lits d'hôpitaux pour 1000 habitants

Le tableau 4 fournit la densité de médecins pour 1000 habitants ainsi que le nombre total de lits d'hôpitaux en densité pour 1000 habitants dans différents pays. Existe-t-il un lien entre ces deux variables ?

Tableau 4 : Médecins en activité et nombre total de lits d'hôpitaux, densité pour 1000 habitants, données pour 2011

Pays	Médecins en activité, densité pour 1000 habitants	Nombre total de lits d'hôpitaux, densité pour 1000 habitants
Autriche	4,83	7,65
Belgique	2,91	6,35
République tchèque	3,64	6,84
Estonie	3,26	5,31
France	3,07	6,32
Allemagne	3,84	8,27
Hongrie	2,95	7,17
Islande	3,51	3,32
Irlande	2,67	2,95
Israël	3,26	3,27
Italie	4,1	3,42
Corée	2,04	9,56
Mexique	2,2	1,68
Nouvelle-Zélande	2,64	2,8
Norvège	3,72	3,32
Pologne	2,19	6,55
Slovénie	2,5	4,63
Espagne	4,1	3,18
Suisse	3,83	4,87
Royaume-Uni	2,81	2,95

Sources : Ressources en santé, Statistiques de l'OCDE sur la santé (base de données), 2013

IV.3. Le tableau de corrélations : Europe 2020

Cet exercice se base sur les données françaises de l'enquête *Eurobaromètre n°79.3* administrée en mai 2013. L'objectif est de déterminer dans quelle mesure les réponses aux sept items de la variable suivante sont liées statistiquement :

« Pour sortir de la crise économique et financière et faire face aux nouveaux défis mondiaux, l'UE a défini une stratégie appelée Europe 2020. Europe 2020 met en avant différentes priorités et objectifs. Parlons-en maintenant.

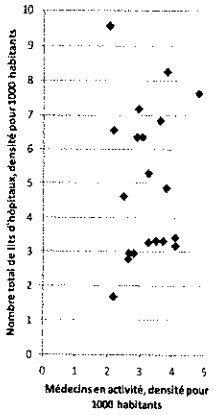
Pour chacune des initiatives suivantes, veuillez me dire dans quelle mesure vous pensez qu'elles sont importantes ou pas pour que l'UE sorte de la crise financière et économique actuelle et se prépare à la prochaine décennie. Veuillez utiliser une échelle de 1 à 10 où '1' signifie "pas du tout importante" et '10' signifie "très importante".

- augmenter l'aide aux politiques de recherche et de développement et transformer les inventions en produits (qb1) ;
- augmenter la qualité et l'attractivité du système d'enseignement supérieur de l'UE (qb2) ;
- développer l'économie en renforçant l'Internet ultrarapide au sein de l'UE (qb3) ;
- soutenir une économie qui utilise moins de ressources naturelles et émet moins de gaz à effet de serre (qb4) ;
- aider la base industrielle de l'UE à devenir plus compétitive par la promotion de l'esprit d'entreprise et par le développement de nouvelles compétences (qb5) ;
- moderniser les marchés de l'emploi en visant l'augmentation du niveau des emplois (qb6) ;
- aider les gens pauvres et socialement exclus et leur permettre de prendre une part active dans la société (qb7). »

1019 personnes ont été interrogées en France pour cette enquête. Interpréter le tableau 5.

.] Le nuage de points associé ne fait pas apparaître un lien linéaire clair entre nsité de médecins en activité et le nombre total de lits d'hôpitaux pour 1000 tants (figure 9).

e 9 : Nuage de points croisant médecins en activité et nombre total de lits d'hôpitaux, densité pour 1000 habitants, données pour 2011



es : Ressources en santé, Statistiques de l'OCDE sur la santé (base de données), 2011

coefficient de corrélation associé vaut 0,04. Il est quasiment nul et montre l'absence de lien de nature linéaire entre ces deux variables.

.] Ici il n'est pas approprié d'utiliser le nuage de points comme première ration du lien entre les items. En effet, il faudrait construire autant de es qu'il y a de liens à mesurer, soit $\frac{7\times6}{2} = 21$ représentations graphiques au . Cela deviendrait vite illisible.

n peut donc directement passer à l'interprétation des coefficients de lation linéaire associés, présentés dans le tableau 5 :

les 21 coefficients de corrélation sont positifs. Les 7 items sont tous liés deux-à-deux de manière positive. Les individus interrogés ont ainsi tendance à se révéler homogènes dans leurs réponses aux différents items, c'est-à-dire qu'ils ont tendance à accorder des notes plutôt élevées deux-à-deux ou des notes plutôt faibles deux-à-deux ;

l'intensité des liens est toutefois variable d'un couple d'items à l'autre : les 21 coefficients sont compris entre 0,19 et 0,53, soit un rapport de 1 à 2,8 entre les coefficients les plus faibles et les plus élevés[1].

mportance donnée à tel ou tel coefficient de corrélation linéaire dépend de la discipline dans quelle il intervient, car les variables employées diffèrent. Ainsi un coefficient de corrélation de 3 ou 0,4 est généralement considéré comme très faible pour un statisticien. Mais en sociologie, n'est pas nécessairement le cas. Tout dépend du contexte d'utilisation de cet outil statistique. effet, nous nous situons ici dans un cas limite (*cf.* tableau 1), puisque les variables analysées nt à la frontière des types ordinal et quantitatif : il s'agit, en effet, d'échelles de 1 à 10

Il faut donc analyser les coefficients dans le détail afin de repérer d'éventuelles logiques thématiques. Les sept items proposés peuvent être classés en trois sujets principaux : les propositions relatives à l'enseignement et à la recherche (y compris la recherche industrielle) dans les items 1 et 2, celles relatives à l'économie et à l'emploi (items 3 à 6) et celle relative au social dans l'item 7. L'item 7 « Aider les gens pauvres et socialement exclus et leur permettre de prendre une part active dans la société » est celui qui se révèle le moins corrélé avec les six autres (coefficients R tous inférieurs à 0,35). Les corrélations les plus fortes apparaissent entre les items 1 « Augmenter l'aide aux politiques de recherche et de développement et transformer les inventions en produits » et 2 « Augmenter la qualité et l'attractivité du système d'enseignement supérieur de l'UE » appartenant au même thème global sur l'enseignement et la recherche (R=0,53). C'est également le cas pour le thème portant sur l'économie et l'emploi, en particulier entre les items 5 « Aider la base industrielle de l'UE à devenir plus compétitive par la promotion de l'esprit d'entreprise et par le développement de nouvelles compétences » et 6 « Moderniser les marchés de l'emploi en visant l'augmentation du niveau des emplois » (R=0,45), 5 et 2 (R=0,43), 5 et 4 « Soutenir une économie qui utilise moins de ressources naturelles et émet moins de gaz à effet de serre » (R=0,41). Enfin, il existe également un lien statistique relativement important entre les items 2 et 3 « Développer l'économie en renforçant l'Internet ultrarapide au sein de l'UE » (R=0,40).

[IV.4.] Le tableau 6 montre que l'âge influe sur le niveau d'intensité de la corrélation entre les items 1 « Augmenter l'aide aux politiques de recherche et de développement et transformer les inventions en produits » et 2 « Augmenter la qualité et l'attractivité du système d'enseignement supérieur de l'UE ». En effet, les quatre coefficients constituent un gradient croissant : plus on est âgé plus le coefficient R est élevé. Autrement écrit, plus on est âgé, plus le lien entre les réponses aux deux items considérés est fort, donc plus on est homogène dans ses réponses à ces deux items (propension à attribuer simultanément soit deux notes élevées, soit deux notes basses). Il faudrait bien entendu calculer ces quatre coefficients pour l'ensemble des 21 associations d'items possibles afin de comprendre si ce phénomène est valable quel que soit le couple d'items considéré.

Cet exercice a mis au jour un élément à prendre en compte lors de l'interprétation d'un coefficient de corrélation : le lien mesuré au niveau global peut varier en fonction de sous-groupes de population. Dans notre exemple, le lien reste positif et varie seulement en intensité. Mais il peut exister des cas où un certain lien au niveau global s'inverse dans certains sous-groupes et pas dans d'autres, d'où l'importance d'être prudent dans l'interprétation de tels coefficients lorsque l'on sait qu'il existe des variations entre sous-groupes.

auxquelles peut s'appliquer une analyse de corrélation. Ceci explique que ces coefficients, tout en étant relativement éloignés de 1, s'avèrent relativement probants pour le sociologue-statisticien (il ne s'agit pas de variables quantitatives pures, pour lesquelles l'éventail des valeurs est important et où chaque unité statistique – *ou presque* – possède une valeur unique).

Les articles d'application conseillés

- Andrew E. Clark, Claudia Senik, « La croissance du PIB rendra-t-elle les habitants des pays en développement plus heureux ? », *Revue d'économie du développement*, vol. 25, n°2, 2011, p. 113-190.
- Hakima Megherbi, Thierry Rocher, Valérie Gyselinck, Bruno Trosseille, Hubert Tardieu, « Évaluation de la compréhension de l'écrit chez l'adulte », *Économie et statistique*, n°424–425, 2009, p. 63-86.

Chapitre 6

La mesure du lien statistique entre deux variables nominales ou ordinales

Nous venons d'étudier dans le chapitre 5 la mesure et l'interprétation du lien entre deux variables quantitatives. Ce chapitre 6 aborde le cas de deux variables qualitatives (nominales ou ordinales). Dans un premier temps, nous allons expliquer comment présenter sous forme de tableau la relation entre deux variables de ce type, puis analyser de manière descriptive leur éventuel lien. Enfin, nous présenterons la procédure du test du Chi-deux[1] qui permet de conclure sur la significativité ou non de ce lien.

Nous nous fonderons sur un exemple tiré de l'enquête TNS Sofres – *TriÉlec* de mars 2012 : l'étude de la relation entre un indicateur de revenus/patrimoine et l'opinion à propos de l'affirmation suivante : « Pour établir la justice sociale, il faudrait prendre aux riches pour donner aux pauvres » (JustSoc)[2].

L'hypothèse sous-jacente, qu'il s'agira de vérifier ici à partir des outils de la statistique bivariée et de l'inférence adaptés au cas de deux variables qualitatives, présume que le niveau socio-économique évalué à partir de la hauteur du revenu et de la dotation patrimoniale influerait sur l'opinion que l'on se fait de l'efficacité de la redistribution pour combattre les inégalités sociales. En d'autres termes, nous souhaitons **expliquer (et s'assurer statistiquement de la légitimité de le faire) l'opinion relative à la justice sociale en fonction de la richesse détenue** (revenus et dotation patrimoniale) par les personnes interrogées dans l'enquête. Cette opinion varie-t-elle dans l'échantillon en fonction du niveau de richesse détenue ? Les écarts observés sont-ils significatifs et peut-on conclure, à partir de cet échantillon, de façon plus globale concernant la population d'intérêt ?

I. La présentation des tableaux croisés

Dans un premier temps, il importe d'apprécier la répartition des personnes interrogées en combinant leurs réponses relatives à leur niveau socio-économique et celles concernant leur opinion vis-à-vis de la redistribution des richesses.

On appelle cette procédure un **tableau croisé** (ou tri croisé) de deux variables nominales. Il s'agit d'un tableau renseignant simultanément les éléments

---

1 Il existe différentes manières d'écrire le Chi-deux que le lecteur ne manquera pas de rencontrer au gré de ses lectures et recherches : Chi-deux, Khi-deux, Chi^2, Khi^2 et enfin la notation mathématique χ^2.

2 Se reporter à l'exercice 2 du chapitre 3 pour la construction de l'indicateur revenus/patrimoine. La seconde variable est quant à elle un des items de la batterie suivante : « Voici maintenant une liste de phrases. Pour chacune d'elles, pouvez-vous me dire si vous êtes ... Pas du tout d'accord / Plutôt pas d'accord / Plutôt d'accord / Tout à fait d'accord ? ».