

De la corrélation à la co-relation

Mesures de dépendance, information et causalité

Hugues Genvrin

Octobre 2026

Résumé

Ce document fait le point sur la notion de corrélation. On en retrace l'histoire, du mot de la langue classique au coefficient de Pearson. On établit les propriétés exactes de ce coefficient et ses limites : il ne mesure que la dépendance linéaire. On présente les autres mesures de dépendance, puis les pièges classiques de l'interprétation. On définit ensuite une *co-relation* rigoureuse fondée sur l'entropie : l'information mutuelle, symétrique, et le coefficient d'incertitude, orienté, qui vaut 1 exactement quand une variable détermine l'autre. On montre comment l'implication logique se traduit en termes probabilistes et entropiques. On explique pourquoi aucune mesure d'association, entropique ou non, ne suffit à établir une causalité, et ce qu'il faut y ajouter. On montre enfin qu'une contrainte de conservation engendre des corrélations sans aucun lien causal entre les termes corrélés : c'est le cas de la troisième loi de Newton comme du modèle d'Ehrenfest.

Table des matières

1	Le mot et son histoire	2
1.1	Dans la langue	2
1.2	Dans les sciences	2
2	Le coefficient de corrélation linéaire	3
3	Autres mesures de dépendance	4
3.1	Coefficients de rangs	4
3.2	Variables qualitatives	4
3.3	Dépendances non linéaires	4
4	Les pièges de l'interprétation	5
4.1	Le quartet d'Anscombe	5
4.2	Les corrélations absurdes	5
4.3	Le paradoxe de Simpson	5
4.4	Sélection et agrégation	5
4.5	Diagnostic et corrélation	6
5	La co-relation informationnelle	6
5.1	Co-relation symétrique : l'information mutuelle	6
5.2	Co-relation orientée : le coefficient d'incertitude	7
5.3	Lien avec les relations et les partitions	7
5.4	L'implication logique	7

6	De la co-relation à la causalité	8
6.1	Ce que mesure une co-relation	8
6.2	Intervenir	9
6.3	Le temps et l'information	9
6.4	Une démarche	9
7	Contraintes et corrélations	10
7.1	Une contrainte de somme engendre des corrélations négatives	10
7.2	Deux exemples physiques	10
7.3	Entropie maximale et corrélations sous contrainte	11
8	Synthèse	11

1 Le mot et son histoire

1.1 Dans la langue

Le mot *corrélation* entre dans le *Dictionnaire de l'Académie française* avec sa deuxième édition, en 1718, avec une définition philosophique : « rapport existant entre deux choses, deux notions, deux faits dont l'un implique l'autre *et réciproquement* ». La réciprocité fait partie de la définition d'origine : la corrélation classique est une implication *mutuelle*. L'Académie illustre aujourd'hui le mot par les termes « père » et « fils », dont l'un ne va pas sans l'autre.

En grammaire, une construction *corrélative* lie deux propositions interdépendantes, dont aucune n'est complète sans l'autre : « *plus* on avance, *plus* la route est longue », « il pleuvait *si* fort *qu'*on ne voyait rien ». Ce sont les exemples linguistiques les plus proches du sens statistique : « plus... plus » exprime exactement une covariation positive.

Il faut les distinguer d'autres constructions :

- « Je pense qu'elle est partie, *puisque* la voiture n'est pas dans le garage » contient une subordonnée circonstancielle de cause, et le raisonnement est une *abduction* : on remonte d'un indice à son explication la plus probable.
- « Si tu avais invité Nathalie, tu aurais passé une meilleure soirée » est un système hypothétique contrefactuel. La forme paratactique, sans subordination, serait : « Tu l'aurais invitée, tu aurais passé une meilleure soirée. »
- « Valérie travaille bien lorsque le temps est gris » n'exprime pas, dans l'usage courant, une implication stricte, mais une *tendance* : Valérie travaille *plus souvent* bien par temps gris. C'est exactement une corrélation au sens statistique, et non une implication logique.

1.2 Dans les sciences

Les moindres carrés. La méthode des moindres carrés est publiée par Legendre en 1805 et justifiée par Gauss en 1809, pour ajuster des orbites à des observations astronomiques entachées d'erreurs ; Boscovich avait proposé dès 1757 une méthode fondée sur la somme des écarts absolus, et Laplace l'avait étudiée. Ces auteurs ajustent des droites et des courbes ; ils ne parlent ni de régression ni de corrélation.

Galton. Francis Galton — savant aux intérêts multiples, à la fois anthropologue, statisticien, explorateur et généticien avant la lettre — observe que les enfants de parents très grands sont en moyenne grands, mais moins que leurs parents. Il nomme ce phénomène *reversion* (1877), puis *regression towards mediocrity* (1886). Dans l'article « Co-relations and their measurement, chiefly from anthropometric data » (*Proceedings of the Royal Society*, 1888), il définit un indice de « co-relation » entre deux mesures du corps. C'est donc la graphie *co-relation* qu'il emploie

d'abord ; la forme soudée *correlation* s'impose ensuite, notamment avec son article « Kinship and correlation » (1890). Son livre *Natural Inheritance* (1889) expose l'ensemble.

Bravais et Pearson. La formule du coefficient apparaît déjà chez Auguste Bravais en 1846, dans une étude des erreurs de position d'un point. Karl Pearson en donne la forme moderne et l'estimateur « produit des moments » en 1896, et développe la corrélation multiple. Yule introduit la corrélation partielle, Spearman (1904) puis Kendall (1938) des coefficients de rangs.

Les mises en garde. Dès 1926, Yule publie « Why do we sometimes get nonsense-correlations between time-series ? », qui montre que deux séries temporelles sans lien peuvent présenter une corrélation forte. Les mises en garde se sont multipliées depuis ; elles font l'objet de la section 4.

2 Le coefficient de corrélation linéaire

Dans toute la suite, X et Y sont des variables aléatoires réelles de variances finies et non nulles.

Définition 2.1. La *covariance* de X et Y est $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$. Le *coefficient de corrélation linéaire* (de Bravais–Pearson) est

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Sur un échantillon $(x_i, y_i)_{1 \leq i \leq n}$, on l'estime par

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}.$$

Théorème 2.2. 1. $-1 \leq \rho(X, Y) \leq 1$, quelle que soit la loi de (X, Y) .

2. $|\rho(X, Y)| = 1$ si et seulement s'il existe des réels a et $b \neq 0$ tels que $Y = a + bX$ presque sûrement ; ρ a alors le signe de b .

3. ρ est invariant par les changements affines croissants de chaque variable : $\rho(a + bX, c + dY) = \rho(X, Y)$ si $bd > 0$.

Démonstration. Posons $U = (X - \mathbb{E}X)/\sigma_X$ et $V = (Y - \mathbb{E}Y)/\sigma_Y$: $\mathbb{E}U^2 = \mathbb{E}V^2 = 1$ et $\rho = \mathbb{E}[UV]$. Pour tout réel λ ,

$$0 \leq \mathbb{E}[(V - \lambda U)^2] = 1 - 2\lambda\rho + \lambda^2.$$

Avec $\lambda = \rho$, on obtient $0 \leq 1 - \rho^2$, d'où le point 1. L'égalité $\rho^2 = 1$ équivaut à $\mathbb{E}[(V - \rho U)^2] = 0$, c'est-à-dire $V = \rho U$ presque sûrement : c'est une relation affine entre X et Y , d'où le point 2. Le point 3 est immédiat, car U et V ne changent pas, ou changent tous deux de signe. $\square$

La normalité des variables n'intervient nulle part : le coefficient est compris entre -1 et 1 pour toute loi. Elle n'intervient que dans la loi d'échantillonnage de r , donc dans les tests et les intervalles de confiance.

Théorème 2.3 (Part de variance expliquée). *La meilleure approximation affine de Y par X , au sens des moindres carrés, est*

$$\hat{Y} = \mathbb{E}Y + \rho \frac{\sigma_Y}{\sigma_X} (X - \mathbb{E}X), \quad \text{et} \quad \min_{a,b} \mathbb{E}[(Y - a - bX)^2] = \sigma_Y^2 (1 - \rho^2).$$

Démonstration. Avec les notations précédentes, minimiser $\mathbb{E}[(Y - a - bX)^2]$ revient à minimiser $\sigma_Y^2 \mathbb{E}[(V - \lambda U)^2] = \sigma_Y^2 (1 - 2\lambda\rho + \lambda^2)$, minimum atteint en $\lambda = \rho$ et égal à $\sigma_Y^2 (1 - \rho^2)$. $\square$

Ainsi ρ^2 est la part de la variance de Y expliquée par une fonction *affine* de X .

Corollaire 2.4 (Régression vers la moyenne). *En variables réduites, la droite de régression de V en U est $\hat{V} = \rho U$. Si $|\rho| < 1$, la valeur prédite est plus proche de la moyenne que la valeur observée.*

C'est exactement le phénomène observé par Galton : la taille des parents et celle des enfants ont une corrélation inférieure à 1, donc les enfants de parents extrêmes « régressent » vers la moyenne. Ce n'est pas un mécanisme biologique, mais une conséquence de $|\rho| < 1$; le même effet se produit dans l'autre sens, des enfants vers les parents.

Proposition 2.5 (Corrélation nulle et indépendance).

1. Si X et Y sont indépendantes, alors $\rho(X, Y) = 0$.
2. La réciproque est fautive : si X est uniforme sur $\{-2, -1, 0, 1, 2\}$ et $Y = X^2$, alors $\rho(X, Y) = 0$, alors que Y est entièrement déterminée par X .
3. Si le couple (X, Y) est gaussien, alors $\rho(X, Y) = 0$ entraîne l'indépendance. Il ne suffit pas que X et Y soient chacune gaussiennes.

Démonstration. 1. L'indépendance donne $\mathbb{E}[XY] = \mathbb{E}X \mathbb{E}Y$. 2. $\mathbb{E}X = 0$ et $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0$ par symétrie, donc $\text{Cov}(X, Y) = 0$. 3. La densité d'un couple gaussien de corrélation nulle est le produit des densités marginales. $\square$

Remarque. Un coefficient nul signifie donc l'absence de dépendance *linéaire*, et non l'absence de dépendance. À l'inverse, $|\rho| = 1$ a un sens exact : une relation affine. Entre les deux, la valeur de ρ dépend à la fois de la force de la dépendance et de sa forme.

3 Autres mesures de dépendance

3.1 Coefficients de rangs

Le coefficient de *Spearman* ρ_S est le coefficient de Pearson calculé sur les rangs des observations. Il vaut 1 si et seulement si Y est une fonction strictement *croissante* de X (sur l'échantillon), qu'elle soit linéaire ou non. Le τ de *Kendall* est la différence entre les proportions de paires concordantes et discordantes. Ces coefficients conviennent aux variables *ordinales* — des grandeurs que l'on sait ranger — et résistent aux valeurs aberrantes. Ils ne conviennent pas aux variables qualitatives sans ordre.

3.2 Variables qualitatives

Pour deux variables catégorielles, on part du tableau de contingence et de la statistique du χ^2 d'indépendance. Pour deux variables binaires, le coefficient φ est le coefficient de Pearson des indicatrices :

$$\varphi = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{p_{1\cdot} p_{0\cdot} p_{\cdot 1} p_{\cdot 0}}}.$$

Pour un tableau $k \times l$, le V de Cramér normalise le χ^2 : $V = \sqrt{\chi^2 / (n (\min(k, l) - 1))} \in [0; 1]$.

3.3 Dépendances non linéaires

- Le *rapport de corrélation* $\eta^2 = \text{Var}(\mathbb{E}[Y \mid X]) / \text{Var}(Y)$ est la part de variance de Y expliquée par une fonction *quelconque* de X . On a toujours $\rho^2 \leq \eta^2 \leq 1$, et $\eta^2 = 1$ si et seulement si Y est une fonction de X .

- La *corrélation maximale* de Gebelein et Rényi est le supremum de $\rho(f(X), g(Y))$ sur toutes les fonctions f et g de variance finie. Elle est nulle si et seulement si X et Y sont indépendantes. Rényi (1959) a montré qu'elle vérifie tous les axiomes qu'il propose pour une mesure de dépendance, ce qui n'est le cas d'aucun des coefficients usuels.
- La *corrélation de distance* de Székely, Rizzo et Bakirov (2007) est elle aussi nulle si et seulement si les variables sont indépendantes, et s'estime facilement.
- L'*information mutuelle*, étudiée à la section 5, possède la même propriété et se relie directement à l'entropie.

La régression *logistique*, souvent citée à ce propos, répond à une autre question : elle modélise la probabilité d'un événement binaire en fonction de variables explicatives. Elle mesure une association, sous forme de rapports de cotes, mais n'est pas une mesure générale de dépendance non linéaire.

4 Les pièges de l'interprétation

4.1 Le quartet d'Anscombe

Anscombe (1973) a construit quatre jeux de onze points qui ont les mêmes moyennes ($\bar{x} = 9$, $\bar{y} = 7,50$), les mêmes variances (11 et 4,12), la même corrélation ($r = 0,816$) et la même droite de régression ($y = 3 + 0,5x$). Pourtant, le premier est un nuage linéaire bruité, le deuxième une courbe parabolique sans bruit, le troisième une droite parfaite perturbée par un point aberrant, et le quatrième une colonne verticale de points complétée par un seul point isolé, qui crée à lui seul toute la corrélation. Un coefficient ne remplace pas le graphique.

4.2 Les corrélations absurdes

Deux marches aléatoires indépendantes ont, sur une période donnée, des corrélations empiriques souvent très éloignées de 0, car chacune a une tendance qui lui est propre. C'est le phénomène décrit par Yule en 1926. Toute série qui croît avec le temps paraît corrélée à toute autre série croissante. Le remède consiste à travailler sur les *variations* plutôt que sur les niveaux, ou à tester la cointégration.

4.3 Le paradoxe de Simpson

Une étude de 1986 sur le traitement des calculs rénaux (Charig et al.) compare deux techniques :

	Petits calculs	Gros calculs	Ensemble
Traitement A	81/87 = 93 %	192/263 = 73 %	273/350 = 78 %
Traitement B	234/270 = 87 %	55/80 = 69 %	289/350 = 83 %

Le traitement A est meilleur dans chaque groupe, mais moins bon sur l'ensemble : il a surtout été donné pour les gros calculs, plus difficiles. La taille du calcul est un *facteur de confusion*, lié à la fois au choix du traitement et au résultat. Les méthodes de stratification, comme celle de Mantel et Haenszel, corrigent cet effet ; mais décider s'il faut stratifier demande de savoir *causalement* quel rôle joue la variable de stratification.

4.4 Sélection et agrégation

- *Biais de Berkson* (1946) : deux maladies indépendantes dans la population peuvent paraître liées parmi les patients hospitalisés, si chacune augmente la probabilité d'hospitalisation. La sélection de l'échantillon crée la corrélation.

- *Erreur écologique* (Robinson, 1950) : une corrélation entre moyennes de groupes, par exemple entre régions, peut être très différente, voire de signe contraire, de la corrélation entre individus.

4.5 Diagnostic et corrélation

En médecine, la *sensibilité* $\Pr(\text{test+} \mid \text{malade})$ et la *spécificité* $\Pr(\text{test-} \mid \text{sain})$ mesurent la qualité d'un test diagnostique, et non une corrélation. La valeur prédictive positive $\Pr(\text{malade} \mid \text{test+})$ dépend en outre de la prévalence, par la formule de Bayes : un test très sensible et très spécifique peut avoir une faible valeur prédictive dans une population où la maladie est rare. C'est encore une asymétrie entre $\Pr(B \mid A)$ et $\Pr(A \mid B)$, que la section suivante retrouve.

5 La co-relation informationnelle

On considère maintenant X et Y à valeurs dans des ensembles finis. On note

$$H(X) = - \sum_x \Pr(X = x) \ln \Pr(X = x)$$

l'entropie de X , $H(X, Y)$ celle du couple et $H(Y \mid X) = H(X, Y) - H(X)$ l'entropie conditionnelle.

5.1 Co-relation symétrique : l'information mutuelle

Définition 5.1. L'*information mutuelle* de X et Y est

$$I(X; Y) = H(X) + H(Y) - H(X, Y) = \sum_{x,y} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} = D(p_{XY} \parallel p_X p_Y).$$

Théorème 5.2. 1. $I(X; Y) = I(Y; X) \geq 0$, avec égalité si et seulement si X et Y sont indépendantes.

2. $I(X; Y) = H(Y) - H(Y \mid X) \leq \min(H(X), H(Y))$.

3. $I(X; Y) = H(Y)$ si et seulement si Y est une fonction de X (presque sûrement).

4. $I(f(X); g(Y)) \leq I(X; Y)$ pour toutes fonctions f et g , avec égalité si f et g sont injectives.

Démonstration. 1. L'entropie relative est positive, et nulle seulement si les deux lois coïncident, c'est-à-dire si $p(x, y) = p(x)p(y)$. 2. Par définition de $H(Y \mid X)$, qui est positive. 3. $H(Y \mid X) = 0$ signifie que, pour chaque x de probabilité non nulle, la loi de Y sachant $X = x$ est concentrée en un point. 4. C'est l'inégalité de traitement des données : transformer une variable ne crée pas d'information. Dans le langage des partitions, $f(X)$ correspond à une partition moins fine, et l'on applique la croissance par raffinement. $\square$

Contrairement au coefficient de Pearson, l'information mutuelle détecte *toute* dépendance, linéaire ou non, et s'applique aux variables qualitatives sans ordre. Elle est invariante par toute recodification bijective des valeurs.

Proposition 5.3 (Cas gaussien). Si (X, Y) est un couple gaussien de corrélation ρ , son information mutuelle, définie à partir des densités, est

$$I(X; Y) = -\frac{1}{2} \ln(1 - \rho^2).$$

Dans le cas gaussien, et seulement dans ce cas, information mutuelle et corrélation linéaire sont donc deux expressions de la même quantité; I tend vers l'infini quand $|\rho| \rightarrow 1$, ce qui traduit une relation déterministe entre variables continues.

5.2 Co-relation orientée : le coefficient d'incertitude

Définition 5.4. Le *coefficient d'incertitude* de Y par rapport à X , dû à Theil, est

$$U(Y | X) = \frac{I(X; Y)}{H(Y)} = 1 - \frac{H(Y | X)}{H(Y)} \in [0; 1].$$

C'est la fraction de l'incertitude sur Y que lève la connaissance de X .

Proposition 5.5. 1. $U(Y | X) = 0$ si et seulement si X et Y sont indépendantes.

2. $U(Y | X) = 1$ si et seulement si Y est une fonction de X .

3. En général, $U(Y | X) \neq U(X | Y)$.

Exemple 5.6. Reprenons X uniforme sur $\{-2, \dots, 2\}$ et $Y = X^2$. On a $\rho = 0$, mais

$$H(X) = \ln 5 \approx 1,609, \quad H(Y) \approx 1,055, \quad I(X; Y) = H(Y) \approx 1,055,$$

$$U(Y | X) = 1, \quad U(X | Y) \approx 0,655.$$

X détermine entièrement Y , mais Y ne détermine pas X : il reste le signe à deviner.

Définition 5.7 (Co-relation). On appelle *co-relation* de X vers Y , notée $X \text{ co}\mathcal{R} Y$, la donnée de la loi jointe de (X, Y) , quantifiée par $U(Y | X)$. On dit que :

- il n'y a *pas de co-relation* si $U(Y | X) = 0$, c'est-à-dire si X et Y sont indépendantes ;
- la co-relation est *totale* de X vers Y si $U(Y | X) = 1$, c'est-à-dire si Y est une fonction de X ;
- la co-relation est *totale réciproque* si $U(Y | X) = U(X | Y) = 1$, c'est-à-dire si X et Y se déterminent mutuellement, par une bijection entre les valeurs prises.

La co-relation totale réciproque correspond exactement à la définition classique de 1718 : deux faits « dont l'un implique l'autre et réciproquement ».

5.3 Lien avec les relations et les partitions

Une variable X définie sur un univers fini Ω muni de la loi uniforme induit une partition de Ω : ses fibres. L'information mutuelle de deux variables est alors l'information mutuelle des deux partitions, et la co-relation totale de X vers Y signifie que la partition de X est *plus fine* que celle de Y . La co-relation totale réciproque signifie que les deux partitions coïncident. Les résultats sur l'entropie des partitions — règle de chaînage, croissance par raffinement, information mutuelle — s'appliquent tels quels.

5.4 L'implication logique

Soient f_1 et f_2 deux énoncés, vus comme des événements de probabilité non nulle. On note $\mathbf{1}_f$ l'indicatrice de f .

Proposition 5.8. Les conditions suivantes sont équivalentes :

1. $f_1 \Rightarrow f_2$ presque sûrement, c'est-à-dire $\Pr(f_1 \wedge \neg f_2) = 0$;
2. $\Pr(f_2 | f_1) = 1$.

En termes d'entropie, $H(\mathbf{1}_{f_2} | f_1 \text{ vrai}) = 0$ — sur la restriction à f_1 , l'énoncé f_2 n'a plus aucune incertitude — si et seulement si $f_1 \Rightarrow f_2$ ou $f_1 \Rightarrow \neg f_2$: l'entropie nulle traduit la détermination, et la valeur prise en donne le sens. L'implication réciproque, $\Pr(f_1 | f_2) = 1$, n'en découle pas. L'équivalence $f_1 \Leftrightarrow f_2$ presque sûre s'écrit $\Pr(f_1 | f_2) = \Pr(f_2 | f_1) = 1$, soit $\Pr(f_1 \Delta f_2) = 0$.

Démonstration. $\Pr(f_2 \mid f_1) = 1 - \Pr(f_1 \wedge \neg f_2) / \Pr(f_1)$, d'où $1 \Leftrightarrow 2$. Sachant f_1 , l'indicatrice de f_2 suit une loi de Bernoulli de paramètre $q = \Pr(f_2 \mid f_1)$, d'entropie nulle si et seulement si $q \in \{0, 1\}$, c'est-à-dire $f_1 \Rightarrow \neg f_2$ ou $f_1 \Rightarrow f_2$. L'exemple ci-dessous montre que la réciproque peut échouer. $\square$

Exemple 5.9. Sur une population, 30 % vérifient f_1 et f_2 , 20 % vérifient f_2 seulement, 50 % ne vérifient aucun des deux. Alors $\Pr(f_2 \mid f_1) = 1$ mais $\Pr(f_1 \mid f_2) = 0,6$. Le coefficient φ vaut 0,655, et les coefficients d'incertitude valent $U(\mathbf{1}_{f_2} \mid \mathbf{1}_{f_1}) \approx 0,396$ et $U(\mathbf{1}_{f_1} \mid \mathbf{1}_{f_2}) \approx 0,449$.

Remarque. Cet exemple montre une subtilité. L'implication est une propriété *locale*, portée par la restriction à f_1 ; les mesures globales, φ comme U , mélangent cette restriction avec le comportement sur $\neg f_1$, où f_2 reste incertain. Ainsi $U(\mathbf{1}_{f_2} \mid \mathbf{1}_{f_1}) < 1$ alors même que $f_1 \Rightarrow f_2$. C'est la probabilité conditionnelle, ou l'entropie sur la restriction à f_1 , et non le coefficient U , qui traduit l'implication. Le coefficient U vaut 1 dans les deux sens exactement quand $\mathbf{1}_{f_2}$ est une fonction de $\mathbf{1}_{f_1}$: $f_2 = f_1$, $f_2 = \neg f_1$, ou f_2 constant.

Remarque (Limites de la traduction probabiliste). — Une probabilité égale à 1 n'est pas une nécessité logique : dans un univers infini, un événement de probabilité 1 peut ne pas se réaliser.

- $\Pr(f_2 \mid f_1)$ n'est pas définie si $\Pr(f_1) = 0$, alors que l'implication matérielle est alors vraie.
- Lewis (1976) a démontré que la probabilité conditionnelle $\Pr(f_2 \mid f_1)$ ne peut pas, en général, être la probabilité d'un énoncé « si f_1 alors f_2 » : la co-relation conditionnelle n'est pas un connecteur logique comme les autres.
- L'implication n'est pas une conjonction : $f_1 \Rightarrow f_2$ équivaut à $\neg f_1 \vee f_2$, et non à $f_1 \wedge f_2$.

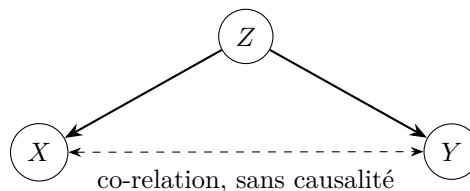
6 De la co-relation à la causalité

6.1 Ce que mesure une co-relation

Toutes les mesures précédentes, y compris les mesures entropiques, sont des fonctions de la seule loi jointe de (X, Y) . Or des structures causales très différentes peuvent produire la même loi jointe.

Proposition 6.1. Soit Z une variable binaire équiprobable, et $X = Z, Y = Z$. Alors $I(X; Y) = \ln 2$ et $U(Y \mid X) = U(X \mid Y) = 1$: la co-relation est totale et réciproque. Pourtant, modifier X de l'extérieur ne change pas Y : X n'a aucun effet sur Y , ni Y sur X .

C'est le schéma d'une *cause commune*. Reichenbach (1956) en a fait un principe : si deux événements sont corrélés, alors soit l'un cause l'autre, soit ils ont une cause commune — en mettant à part les corrélations créées par la sélection de l'échantillon (biais de Berkson) et par les contraintes (section 7).



En particulier, la condition « l'effet n'a qu'un antécédent », ou $H(\text{cause} \mid \text{effet}) = 0$, dit que l'on peut *remonter* de l'effet à la cause : c'est une propriété de *diagnostic*, utile, mais qui ne dit pas que la cause produit l'effet. Une restriction mal choisie du domaine des causes, qui oublie une variable comme Z , peut créer ou masquer une dépendance : c'est le paradoxe de Simpson.

6.2 Intervenir

La causalité porte sur ce qui se passe quand on *agit* sur une variable. Pearl la formalise par l'opérateur do : $\Pr(Y \mid \text{do}(X = x))$ est la loi de Y quand on impose $X = x$ de l'extérieur, en coupant les causes habituelles de X . En général,

$$\Pr(Y \mid \text{do}(X = x)) \neq \Pr(Y \mid X = x).$$

Dans la proposition 6.1, $\Pr(Y = 1 \mid X = 1) = 1$ mais $\Pr(Y = 1 \mid \text{do}(X = 1)) = \frac{1}{2}$.

Théorème 6.2 (Randomisation). *Si la valeur de X est tirée au sort, indépendamment de tous les autres facteurs, alors $\Pr(Y \mid \text{do}(X = x)) = \Pr(Y \mid X = x)$: la co-relation observée est causale.*

Idée de la preuve. Dans le langage des résultats potentiels de Rubin, notons $Y(x)$ la valeur que prendrait Y si l'on imposait $X = x$. Le tirage au sort rend X indépendante de la famille $(Y(x))_x$. Alors $\Pr(Y = y \mid X = x) = \Pr(Y(x) = y \mid X = x) = \Pr(Y(x) = y)$, qui est la loi sous intervention. $\square$

C'est pourquoi l'essai randomisé reste la référence pour établir un effet causal. En l'absence de randomisation, trois voies existent :

- rendre explicites les hypothèses causales par un *graphe* et identifier les variables à ajuster (critère de la porte arrière de Pearl) ;
- exploiter des expériences naturelles ou des variables instrumentales ;
- à défaut, rassembler des arguments convergents : ce sont les critères de Bradford Hill (1965), parmi lesquels la force et la constance de l'association, la temporalité, le gradient dose-effet et la plausibilité.

6.3 Le temps et l'information

Les mesures orientées dans le temps — causalité au sens de Granger (1969), entropie de transfert de Schreiber (2000) — demandent si le passé de X améliore la prédiction de Y au-delà du passé de Y . L'entropie de transfert s'écrit comme une information mutuelle conditionnelle :

$$T_{X \rightarrow Y} = I(Y_{t+1}; X_t^{(k)} \mid Y_t^{(l)}).$$

Ces mesures ont une direction, ce qui les rapproche de la causalité, mais elles restent des mesures de *prédiction* : une cause commune qui agit sur X avant d'agir sur Y suffit à produire $T_{X \rightarrow Y} > 0$.

6.4 Une démarche

La démarche suivante place chaque outil à sa place :

1. *Délimiter le système* : quelles variables, quelle population, quelle échelle de temps ; dessiner un graphe causal provisoire.
2. *Décrire les associations* : graphiques d'abord, puis coefficients adaptés au type des variables — Pearson, rangs, φ ou V de Cramér, information mutuelle.
3. *Évaluer la détermination* : les coefficients d'incertitude $U(Y \mid X)$ et $U(X \mid Y)$ indiquent dans quel sens, et à quel degré, une variable détermine l'autre.
4. *Contrôler les facteurs de confusion* : stratification, ajustement, en s'appuyant sur le graphe.
5. *Conclure à la causalité* seulement par randomisation, par une stratégie d'identification explicite, ou par un faisceau d'arguments de type Bradford Hill.

7 Contraintes et corrélations

7.1 Une contrainte de somme engendre des corrélations négatives

Théorème 7.1. Soient $X_1, \dots, X_k$ des variables échangeables, de variance $\sigma^2 > 0$, dont la somme est constante : $X_1 + \dots + X_k = c$. Alors, pour $i \neq j$,

$$\rho(X_i, X_j) = -\frac{1}{k-1}.$$

Démonstration. Notons ρ la corrélation commune, identique pour toutes les paires par échangeabilité. La somme étant constante,

$$0 = \text{Var} \left(\sum_i X_i \right) = k\sigma^2 + k(k-1)\rho\sigma^2,$$

d'où $\rho = -1/(k-1)$. □

Proposition 7.2 (Loi multinomiale). Si $(N_1, \dots, N_k)$ suit une loi multinomiale de paramètres n et $(p_1, \dots, p_k)$, alors

$$\text{Cov}(N_i, N_j) = -np_i p_j, \quad \rho(N_i, N_j) = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}} \quad (i \neq j).$$

Démonstration. On écrit $N_i = \sum_{m=1}^n \mathbf{1}_{\{\text{tirage } m \text{ dans } i\}}$. Pour un même tirage, les indicatrices de i et de j ne sont jamais vraies ensemble, donc leur covariance vaut $0 - p_i p_j$; des tirages différents sont indépendants. Ainsi $\text{Cov}(N_i, N_j) = -np_i p_j$, et $\text{Var } N_i = np_i(1-p_i)$. □

Dans les deux cas, les termes corrélés n'agissent pas les uns sur les autres : c'est la contrainte commune qui les lie. Une contrainte joue le rôle d'une cause commune.

7.2 Deux exemples physiques

Le modèle d'Ehrenfest. n boules se répartissent entre deux urnes, avec K boules dans la première et $n-K$ dans la seconde. Les deux effectifs ont une corrélation exactement égale à -1 (théorème 7.1 avec $k=2$, ou théorème 2.2, puisque $n-K$ est une fonction affine décroissante de K). Aucune urne ne « cause » l'effectif de l'autre.

La troisième loi de Newton. « L'action est toujours égale et opposée à la réaction » :

$$\vec{F}_{1/2} = -\vec{F}_{2/1}.$$

Si l'on observe un grand nombre d'interactions entre deux corps, chaque composante de $\vec{F}_{1/2}$ est exactement l'opposée de la composante correspondante de $\vec{F}_{2/1}$: leur corrélation vaut -1 . Mais cette égalité n'est pas un fait statistique observé ; c'est une *loi*, équivalente, pour un système isolé de deux corps, à la conservation de la quantité de mouvement $\vec{p}_1 + \vec{p}_2$. La corrélation -1 est la trace statistique de cette contrainte de conservation. On ne dira pas que la force exercée par le corps 1 « cause » la force exercée par le corps 2 : les deux sont les deux faces d'une même interaction.

7.3 Entropie maximale et corrélations sous contrainte

Le principe d'entropie maximale de Jaynes (1957) donne la forme générale de ce phénomène. Parmi toutes les lois qui respectent des contraintes données, on choisit celle d'entropie maximale : c'est la loi qui n'ajoute aucune information au-delà des contraintes. Sans contrainte, la loi d'entropie maximale sur un produit d'ensembles finis est la loi uniforme, sous laquelle les coordonnées sont indépendantes. Toute corrélation qui apparaît dans la loi d'entropie maximale est donc *entièrement due aux contraintes*. Le modèle d'Ehrenfest l'illustre : la loi stationnaire des micro-états est uniforme sur les 2^n répartitions, mais la contrainte « n boules au total » suffit à lier les effectifs des deux urnes.

Ce cadre donne un sens précis à l'idée d'une co-relation « sous contrainte entropique » : connaissant les contraintes d'un système, on peut *prévoir*, avant toute observation, les corrélations qu'elles imposent. Une corrélation observée qui va au-delà de ces corrélations prévues signale un lien supplémentaire — causal, de cause commune ou de sélection — qu'il reste à identifier.

8 Synthèse

Mesure	Vaut 0 quand	Valeur extrême
Pearson ρ	pas de dépendance linéaire	$ \rho = 1$: relation affine
Spearman, Kendall	pas de tendance monotone d'ensemble	1 : relation strictement croissante
φ , V de Cramér	indépendance (qualitatives)	1 : association parfaite
Rapport de corrélation η^2	$\mathbb{E}[Y X]$ constante	1 : Y fonction de X
Information mutuelle I	indépendance	$\min(H(X), H(Y))$
Coefficient $U(Y X)$	indépendance	1 : Y fonction de X (orienté)
$\Pr(f_2 f_1)$	—	1 : implication $f_1 \Rightarrow f_2$

Conclusion. La corrélation de Pearson n'est qu'une des mesures possibles de la co-relation : elle ne voit que la dépendance linéaire et elle est symétrique. L'entropie fournit une co-relation plus générale, l'information mutuelle, qui détecte toute dépendance, et une co-relation orientée, le coefficient d'incertitude, qui distingue le sens de la détermination ; l'implication logique s'y lit comme une détermination sur une restriction. Aucune de ces mesures ne suffit à établir une causalité : il faut une intervention, ou des hypothèses causales explicites. Enfin, les contraintes de conservation produisent à elles seules des corrélations exactes, sans causalité entre les termes : c'est ce que montrent la troisième loi de Newton, le modèle d'Ehrenfest et le principe d'entropie maximale.

Références

- [1] F. J. Anscombe, « Graphs in statistical analysis », *The American Statistician* 27, 1973.
- [2] J. Berkson, « Limitations of the application of fourfold table analysis to hospital data », *Biometrics Bulletin* 2, 1946.
- [3] A. Bravais, « Analyse mathématique sur les probabilités des erreurs de situation d'un point », *Mémoires présentés par divers savants à l'Académie royale des sciences* 9, 1846.

- [4] C. R. Charig, D. R. Webb, S. R. Payne, J. E. Wickham, « Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy », *British Medical Journal* 292, 1986.
- [5] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley, 2^e éd., 2006.
- [6] F. Galton, « Regression towards mediocrity in hereditary stature », *Journal of the Anthropological Institute* 15, 1886.
- [7] F. Galton, « Co-relations and their measurement, chiefly from anthropometric data », *Proceedings of the Royal Society of London* 45, 1888.
- [8] F. Galton, *Natural Inheritance*, Macmillan, 1889.
- [9] C. W. J. Granger, « Investigating causal relations by econometric models and cross-spectral methods », *Econometrica* 37, 1969.
- [10] A. B. Hill, « The environment and disease : association or causation? », *Proceedings of the Royal Society of Medicine* 58, 1965.
- [11] E. T. Jaynes, « Information theory and statistical mechanics », *Physical Review* 106, 1957.
- [12] M. G. Kendall, « A new measure of rank correlation », *Biometrika* 30, 1938.
- [13] D. Lewis, « Probabilities of conditionals and conditional probabilities », *The Philosophical Review* 85, 1976.
- [14] K. Pearson, « Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia », *Philosophical Transactions of the Royal Society A* 187, 1896.
- [15] J. Pearl, *Causality*, Cambridge University Press, 2^e éd., 2009.
- [16] H. Reichenbach, *The Direction of Time*, University of California Press, 1956.
- [17] A. Rényi, « On measures of dependence », *Acta Mathematica Academiae Scientiarum Hungaricae* 10, 1959.
- [18] W. S. Robinson, « Ecological correlations and the behavior of individuals », *American Sociological Review* 15, 1950.
- [19] D. B. Rubin, « Estimating causal effects of treatments in randomized and nonrandomized studies », *Journal of Educational Psychology* 66, 1974.
- [20] T. Schreiber, « Measuring information transfer », *Physical Review Letters* 85, 2000.
- [21] C. Spearman, « The proof and measurement of association between two things », *American Journal of Psychology* 15, 1904.
- [22] G. J. Székely, M. L. Rizzo, N. K. Bakirov, « Measuring and testing dependence by correlation of distances », *Annals of Statistics* 35, 2007.
- [23] G. U. Yule, « Why do we sometimes get nonsense-correlations between time-series? », *Journal of the Royal Statistical Society* 89, 1926.