Jean-Jacques Droesbeke est professeur à l'Université Libre de Bruxelles.
Philippe Tassi a été DGA de Médiamétrie ; il est membre de la Commission Nationale des Sondages.

Introduction

La notion de corrélation est connue de tous les statisticiens de formation, bien sûr, mais aussi de nombreux spécialistes d'autres domaines utilisant la statistique comme outil d'analyse et de synthèse. Le célèbre *coefficient de corrélation linéaire* entre deux variables numériques X et Y est généralement désigné par la lettre *r* quand on possède une série de *n* observations simultanées $\{(x_i, y_i), i = 1,2 \dots n\}$ de ces variables ou par ρ quand il s'agit d'étudier un couple de variables aléatoires dont la distribution de probabilité caractérise une population. Le coefficient *r* est défini comme étant le rapport entre la *covariance* $s_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})\,(y_i - \bar{y})$ – où $\bar{x}\ et\ \bar{y}$ sont les moyennes arithmétiques respectives des valeurs observées pour X et Y – et le produit des écarts-types $s_x = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$ et $s_y = \sqrt{\frac{1}{n}\sum_i (y_i - \bar{y})^2}$. Cette définition fête ses 130 ans en 2026 (Pearson, 1896). Ce concept a été étendu assez rapidement à plus de deux variables, complété par les notions de *corrélation totale, corrélation partielle*, *corrélation de rangs*… avant d'être un outil clé de l'économétrie et des séries temporelles, notamment par les modèles ARIMA (*AutoRegressive Integrated Moving Average*).

Le coefficient *r* est destiné à mesurer l'intensité d'une *association* entre X et Y, sans nécessairement chercher à déterminer une *relation de causalité*. « Ce n'est pas parce que vous marchez à côté de quelqu'un dans la rue que vous l'accompagnez » (Droesbeke et Vermandele, 2016) ; si l'usage d'un coefficient de corrélation est maintenant considéré tout à fait naturel, son histoire ne s'est pas faite en un jour et mérite d'être connue.

Avant des formules, des mots

Relatio existe en latin classique (Cicéron, Quintilien, Sénèque, …). Mais *correlatio* est rattaché à un latin bien postérieur ; la première référence est attribuée à Beda Venerabilis (vers 672 - 736), moine anglais, historien et scientifique, auteur de *De Natura Rerum* et *De Arte Metrica*.

Le Grand Larousse du XX^{e} siècle (1929) définit la corrélation comme « le rapport d'objets ou de termes dont l'un appelle logiquement l'autre » ; dans la rubrique « encyclopédie », deux exemples sont le lien entre les différentes parties d'un organisme et la corrélation entre les êtres vivants d'un même milieu, en référence à Darwin. Par contre, le « Grand Larousse Universel » de 1989 y consacre une demi-page, la définition première étant « la dépendance ou liaison apparente entre deux choses, deux termes » ; dans deux rubriques (Probabilités et Statistique), il donne explicitement la formule classique présentée ci-dessus, citant même les noms de ses auteurs, Auguste Bravais (1811-1863) – dans un article consacré à la distribution normale bivariée en 1844 – et Karl Pearson (1857-1936) – dans son article de 1896.

La causalité précède la corrélation

La notion de causalité a été formalisée via la philosophie par des personnalités comme Platon (env. -428 à -347) ou Aristote (-385 à -322). « Tout ce qui naît, naît nécessairement par l'action d'une cause », affirmait Platon. Aristote, dans sa *Physique*, établit un lien de dépendance entre connaissance et causes défini selon plusieurs types, dont le *hasard.*

Il faut évoquer aussi d'autres philosophes comme David Hume (1711-1776), auteur de *L'enquête sur l'entendement humain* (Hume, 1748). Pour lui, la relation de cause à effet se caractérise par des propriétés simples : causes et effets se touchent dans l'espace et se suivent immédiatement dans le temps. La perception est au centre de son discours, conduisant à des expériences, sources de nos connaissances. Hume définit deux types de perceptions : celles liées à nos sens, et les idées, plus abstraites mais plus maniables, surtout si elles sont simples. L'avantage de ces dernières est de pouvoir être utilisées par l'« entendement » pour créer des idées complexes. Il existe pour cela trois relations privilégiées : la *ressemblance*, la *contiguïté* et la *causalité*.

La causalité doit beaucoup au développement des sciences. Dans ce domaine, le rôle de Francis Bacon (1561-1626) fut prépondérant. Pour lui, l'expérimentation est un instrument essentiel pour comprendre le processus de causalité. Plus tard, Pierre Simon de Laplace (1749-1827) puise dans les travaux des mathématiciens du XVIII^{e} siècle l'idée de décrire le hasard pour réaliser des estimations précises (distances entre objets célestes…). À la même époque, Carl Friedrich Gauss (1777-1855) relie l'usage d'une moyenne arithmétique comme milieu d'un ensemble de valeurs et le recours au critère des moindres carrés — dont il revendique la paternité en dépit des travaux d'Adrien-Marie Legendre (1752-1833) — au rôle central d'une loi des erreurs particulière qualifiée de « normale » par Francis Galton (1822-1911). De son côté, Laplace met cette distribution des erreurs d'observation à l'honneur en recourant à un théorème qualifié par la suite de « limite central » pour introduire les concepts d'*estimation* et de *précision* approfondis, un siècle plus tard, par des statisticiens anglais dont nous reparlerons (Droesbeke, Tassi, 2015).

« Appliquons aux sciences politiques et morales la méthode fondée sur l'observation et le calcul, cette méthode qui nous a si bien servi dans les sciences naturelles », écrit Laplace en 1814. Un de ses admirateurs, le belge Adolphe Quetelet (1796-1874), s'inspire de ses réflexions en astronomie pour les appliquer à l'étude des caractéristiques physiques et morales des populations (Droesbeke, 2021). Naît la notion d'*homme moyen* qui animera les débats entre statisticiens durant tout le XIX^{e} siècle (Armatte, Droesbeke, 2023). Ses écrits contiennent de nombreuses idées sur des dépendances entre variables, comme l'influence de l'instruction sur le nombre de crimes, dans un mémoire publié en 1831.

Réversion, régression et puis corrélation

Francis Galton est un touche-à-tout très cultivé, passionné d'anthropologie et de géographie. En 1859, il réoriente ses centres d'intérêt vers l'hérédité, pièce centrale du mécanisme que son cousin Charles Darwin (1809-1882) lui a révélé. *Hereditary Genius* est son ouvrage-clé (Galton, 1869).

Si l'*homme moyen* est, pour Quetelet, le porteur de tout, les variations autour de cette moyenne ne sont pour lui que le résultat d'influences variables peu intéressantes. Pour Galton, au contraire, ce qui compte dans la description des attributs humains, ce sont les écarts à la moyenne, déviations lui permettant de développer ses réflexions (Droesbeke *et al.*, 2006). Étudiant la distribution des poids des graines de pois de senteur sur deux générations, il constate que le poids médian des graines « enfant » est une fonction *linéaire* du poids des graines « parent » avec une pente inférieure à 1, et donc le poids médian de la descendance dévie moins de la médiane « population » que le poids des parents : il va l'appeler d'abord *réversion* puis *régression* en 1885. La même publication fait état d'un autre ensemble de données, devenu célèbre, sur l'étude des tailles de parents et d'enfants. Il introduit un *coefficient de réversion r* devenu définitivement *coefficient de corrélation* – on ajoute parfois « linéaire » – sous la plume de Pearson. L'étude de l'association entre frères ou entre les longueurs des bras ou des jambes d'êtres humains ou d'animaux va le conduire à des *co-relations* qui deviendront vite des *corrélations* (1888).

La fin du XIX^{e} siècle et les apports d'Edgeworth, Pearson et Yule

Trois acteurs importants de l'histoire occupent la fin du siècle : Francis Edgeworth (1845-1926), Pearson et George Udny Yule (1871–1951). Ils se sont croisés et recroisés sur les mêmes domaines, la personnalité de Pearson créant des conflits de concurrence.

Porté vers l'économie mathématique, l'Irlandais Edgeworth est attiré par la statistique mathématique sans participer à l'engagement de l'époque pour l'eugénisme. Contacté par Galton, Edgeworth perçoit très tôt l'intérêt de son travail, ce qui lui permet de formaliser, en 1885, les principes de l'*analyse de la variance* avant que Ronald Fisher (1890-1962) ne développe cette méthode. En 1893, Edgeworth exprime le coefficient *r* sous une forme qualifiée de moment-produit, somme normée de produits des valeurs des deux variables considérées, trois ans avant que Pearson fasse la même proposition en 1896.

Le britannique Karl Pearson est d'abord spécialiste de la littérature allemande, puis passe au droit et, en 1885, aux mathématiques appliquées. Il devient un mathématicien statisticien innovant, connu surtout pour son test du χ^2. Cumulant une forte expertise scientifique et des talents d'entrepreneur, il ajoute à ses compétences statistiques une grande capacité d'organisation : il est un vrai manager (Desrosières, 1993).

Dans la lignée de Galton, Pearson s'intéresse à la génétique et à la biométrie. Dans le contexte de l'eugénisme, il présente, en 1896, le coefficient *r* sous l'hypothèse de normalité, « norme » en biologie ; sous cette même loi, il crée les corrélations multiple et partielle. Il faut dire que Edgeworth est un mathématicien discret alors que Pearson a un caractère plus « affirmé ».

Yule est un mathématicien écossais. La corrélation est au centre de ses préoccupations dès 1895, mais il ne recourt pas à la loi normale, contrairement à Pearson (Yule, 1895) ; cela provoque quelques « frictions » venant de celui-ci. Yule introduit aussi les notions de corrélations multiple et partielle, mais dans un contexte non gaussien. En 1900, inspiré par des travaux de Quetelet, il crée un *coefficient d'association* pour la mesure de « l'influence des causes » – noté Q en son

hommage – dans le cadre d'une partition selon une table de contingence à deux lignes et deux colonnes :

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad Q = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11}a_{22} + a_{12}a_{21}}$$

Le début du XX^{e} siècle

Dans la continuité de la fin du siècle précédent, Pearson et Yule poursuivent leurs contributions, s'opposant sur le rôle de la normalité et ses conséquences. Pearson et ses élèves usent d'une dialectique assez violente (Droesbeke, Tassi, 1990). Yule se retire de cette confrontation ; ses hommages officiels ou privés lors du décès de Pearson, en 1938, montrent la dimension humaine et l'ouverture du personnage.

Si l'hérédité, au sens large, a été en lien avec le développement initial de la notion de corrélation, une autre thématique y contribue à la charnière des années 1900 : la psychométrie. Galton s'y était déjà intéressé pour établir une échelle de mesure des aptitudes humaines (Galton, 1869).

Les Français Alfred Binet (1857 – 1911) et Théodore Simon (1873 – 1961) sont mondialement reconnus comme les initiateurs de la mesure en psychologie via des « tests psychométriques » (Binet, 1905 ; Binet, Simon, 1905) pour détecter des enfants en retard, dans un but uniquement pédagogique. Leurs travaux sont le départ d'extensions multiples, notamment aux États-Unis. À l'université de Stanford, ils conduisent à une logique très différente avec la construction pour tous d'un coefficient QI (quotient intellectuel).

Ce nouveau type de données permet à Charles Spearman (1863 – 1945), psychologue anglais, mathématicien et statisticien reconnu à l'*University College* de Londres, de mesurer la corrélation, non pas entre les valeurs $\{(x_i, y_i); i = 1,2, \dots, n\}$ d'un couple de variables (X, Y), mais entre leurs positionnements sur des échelles associées aux questions des tests, lorsqu'on les classe dans l'ordre croissant, donc aux *rangs* engendrés. Ainsi est né le *coefficient de corrélation de rangs* de Spearman.

Plus généralement, à partir de 1904, travaillant sur l'intelligence, Spearman la quantifie par le « facteur g d'intelligence » avec une publication intitulée « théorie des deux facteurs » (Spearman, 1914). L'analyse des corrélations entre les résultats des tests et leurs composantes fait ressortir deux « causes » d'intelligence, appelées facteurs. Le premier est général, noté « facteur g » (« intelligence générale »), le second désigne les « facteurs spécifiques » (« facteur s »). En 1904, Spearman porte son attention sur *g,* avant que *s* ne prenne de l'importance pour sa publication de 1914. Spearman est donc à l'origine de l'*analyse factorielle généralisée*, ainsi nommée en 1931 par Louis Thurstone (1887-1955).

La première partie du XX^{e} est également marquée par de nombreuses avancées. Mentionnons-en quatre :

a) Si trois variables (X, Y, Z) sont observées, le *coefficient de corrélation partielle* entre X et Y est le coefficient de corrélation linéaire appliqué aux parties de X et Y « non expliquées par Z » (Yule, 1897 ; Fisher, 1921, 1924).

b) Comme Spearman, le statisticien anglais Maurice Kendall (1907 – 1983) conçoit, en 1938, un indicateur fondé sur les rangs (Kendall, 1938). À partir des valeurs $\{(x_i, y_i); i = 1,2, \dots, n\}$, on considère tous les couples $\{(x_i, y_i), (x_j, y_j); i, j = 1,2, \dots, n\}$. Un couple est *concordant* ou *cohérent* si ($x_i < x_j$ et $y_i < y_j$), ou si ($x_i > x_j$ et $y_i > y_j$) ; il est *discordant* ou i*ncohérent* si ($x_i < x_j$ et $y_i > y_j$) ou si ($x_i > x_j$ et $y_i < y_j$). Le *coefficient τ de Kendall*, est défini par : $\tau = [N(C) - N(I)]/[n(n-1)/2]$, où $N(C)$ est le nombre de couples cohérents et $N(I)$ celui des couples incohérents.

c) Herman Wold (1908 – 1992), statisticien-économiste suédois d'origine norvégienne, publie en 1938 un théorème à l'origine des futurs modèles autorégressifs (AR) et *moving-average* (MA). Le théorème (de décomposition) de Wold est majeur dans l'analyse des séries temporelles et des processus stochastiques. Au passage, il introduit les concepts d'autorégression, d'autocorrélation et de corrélogramme (Wold, 1938).

d) Le coefficient de détermination, noté R^2 ou r^2, bien connu des économètres, est dû à Pearson ; il est égal au pourcentage de la variance de la variable dépendante expliquée par les variables explicatives, et mesure la qualité d'une régression linéaire.

Conclusion

L'histoire de la corrélation se situe dans des contextes scientifiques et philosophiques majeurs. La statistique a longtemps évacué le problème de la causalité considérant qu'il relevait des domaines d'application et des théories afférentes. Une très large palette de modèles et méthodes pour l'analyse causale s'est peu à peu constituée depuis le début des années 1980, peu répandue en dehors d'un cercle assez restreint. En revanche, la corrélation est un concept bien connu et très utile si on prend soin de préciser à chaque fois son interprétation historique. N'oublions pas que l'on considère, encore trop souvent, une forte corrélation comme synonyme de causalité… même en intelligence artificielle.

Éléments de bibliographie

Armatte M., Droesbeke J-J., *Quetelet. L'œuvre probabiliste (1828-1874)*, Paris, Editions de l'Institut des Études Démographiques, Coll. Les Classiques de l'Économie, 2023.

Binet A., *À propos de la mesure de l'intelligence*, L'année psychologique, 1905

Binet A., Simon Th., *Sur la nécessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence*, L'année psychologique, 1905

Bravais A., *Analyse mathématique sur les probabilités des erreurs de situation d'un point,* Mémoires de l'Institut de France, 9, 1844.

Desrosières A., *La politique des grands nombres*, Ed. La Découverte, Paris, 1993

Droesbeke J-J., *Adolphe Quetelet (1796-1874). Passeur d'idées*, Bruxelles, Académie royale de Belgique, 2021.

Droesbeke J-J., Lejeune M., Saporta G., *La corrélation et ses dérivés : le rôle de Galton dans leur histoire,* in Droesbeke J.-J., Lejeune M. et Saporta G. (éds), *Analyse statistique des données spatiales*, Paris, Technip, 1-15, 2006.

Droesbeke J-J., Tassi Ph., *George Udny Yule ou comment (ne pas) parler de corrélation*, *Statistique et Analyse des Données*, Vol. 15, n°1, 1990, …

Droesbeke J-J., Tassi Ph., *Histoire de la Statistique* (2ème éd.), Presses Universitaires de France, Collection *Que Sais-je* ?, 1997 (2015).

Droesbeke J-J., Vermandele C., *Les nombres au quotidien. Leur histoire, leurs usages*, Paris, Technip, 2016.

Fisher R. A., *On the « probable error » of a coefficient of correlation deduced from a small sample*, Metron 1, 1921

Fisher R. A., *The Distribution of the Partial Correlation Coefficient*, Metron 3, 1924

Galton F., *Hereditary Genius* : *An Inquiry Into Its Laws and Consequences*, MacMillan, 1869.

Galton F., *Co-relations and their measurement, chiefly from anthropometric data*, *Proceedings of the Royal Society of London*, 45, 1888, 135-145.

Hume D., *Enquiry Concerning Human Understanding* (*Philosophical Essays Concerning Human Understanding*), traduction française *Enquête sur l'entendement humain*, Millar Andrew (Londres), 1748, puis Cooper M. (Londres (1751)

Kendall M., *A new measure of rank correlation*, Biometrika, Vol. 30, n°1-2, 1938

Laplace Pierre Simon (de), *Essai philosophique sur les probabilités,* Paris, Veuve Courcier, 1814 (5e édition revue et augmentée, 1825, rééditée en 1986, avec une préface de R. Thom et une postface de B. Bru, Paris, C. Bourgois).

Pearson K., *Mathematical Contributions to the Theory of Evolution – III – Regression, Heredity and Panmixia*, Philosophical Transactions of the Royal Society of London A, 187, 1896, 253-318.

Quetelet A., *Recherches sur le penchant au crime aux différents âges*, Nouveaux mémoires de l'Académie royale des sciences et belles-lettres de Bruxelles, Tome 7, 1831, 87 pages.

Spearman C., *The theory of two factors*, Psychological Review, n°21, 1914

Stigler S., *The History of Statistics, the Measurement of uncertainty before 1900*, The Belknap Press of Harvard University Press, Cambridge (USA), 1986

Thurstone L., *Multiple Factor Analysis*, Psychological Review, n°38, 1931

Wold H., *A Study in the Analysis of Stationary Time Series*, Almqvist & Wiksell, Uppsala, 1938

Yule G. U., *On the Correlation of Total Pauperism with Proportion of Outrelief*, Economic Journal, 5, 1895

Yule G. U., *On the Significance of Bravais' Formulae for Regression,* Proceedings Royal Society, Ser. A, Londres, Vol. 60, 1897