JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

M.-V. MUSHAM

Sur l'interprétation du coefficient de corrélation

Journal de la société statistique de Paris, tome 88 (1947), p. 134-138

<http://www.numdam.org/item?id=JSFS_1947__88__134_0>

Sur l'interprétation du coefficient de corrélation

En 1846, Auguste Bravais publia dans les *Mémoires de l'Institut de France* une communication intitulée : « Analyse mathématique sur les probabilités des erreurs de situation d'un point ». C'est dans ce mémoire que les idées et les formules qui forment la base de la théorie de la corrélation,, furent exposées pour la première fois. La présente étude est dédiée au centième anniversaire de cet événement.

Pendant le siècle qui s'est écoulé depuis qu'Auguste Bravais présenta son célèbre mémoire sur l'analyse mathématique sur les probabilités des erreurs de situation d'un point, la théorie de la corrélation est devenue un des éléments les plus importants de la méthode statistique. La théorie a été fondée solidement du point de vue logique, la technique a été raffinée par la méthode des corrélations partielles, et le problème de la distribution des erreurs du coefficient de corrélation a été résolu élégamment.

Mais, d'autre part, un des problèmes les plus fondamentaux du calcul des corrélations n'a pas trouvé de solution simple et générale : c'est le problème de l'interprétation de la valeur numérique du coefficient de corrélation de Pearson r. Certains auteurs considèrent la fraction qu'on obtient à l'aide de la formule du moment des produits de Bravais, multipliée par 100, comme un pourcentage; d'autres prétendent que le carré du coefficient de corrélation peut être interprété comme un pourcentage; et souvent l'idée d'assimiler le coefficient de corrélation à un pourcentage est rejetée *grosso modo*. Il serait donc important de clarifier les conditions théoriques dans lesquelles l'interprétation du coefficient de corrélation ou de son carré comme un pourcentage est admissible et de définir les types correspondants d'applications pratiques.

Considérons de prime abord les cas où deux mesures sont faites sur des individus appartenant à deux différentes espèces : chez les individus d'une espèces les deux mesures sont liées par une fonction linéaire et, chez ceux de l'autre espèce, elles sont stochastiquement indépendantes. Il est évident que dans ce cas le coefficient de corrélation pour tous les individus est $r = \dfrac{m}{m+n}$, où m et n représentent respectivement le nombre d'individus des deux espèces. Le coefficient de corrélation montre dans ce cas le pourcentage des « véritables » paires parmi toutes les paires de mesures.

Dans un cas plus général, m paires de mesures seront corrélées avec un coefficient r_m et n paires avec un coefficient r_n. Le coefficient de corrélation pour toutes les paires sera alors la moyenne pondérée de r_m et r_n,

$$r = \frac{m\, r_m + n\, r_n}{m+n}$$

La dernière formule pourra être appliquée par exemple à l'étude des jumeaux (5). En effet, tout échantillon de jumeaux contient des couples monovitellins et hétérovitellins, et la corrélation entre les mesures d'une qualité

quelconque (poids, hauteur, intelligence, etc...) sera différente pour chacun des deux types. La formule donnée plus haut permettra de calculer une des inconnues r_m, r_n et $\frac{m}{n}$, si r a été déterminé expérimentalement et si des hypothèses ont été faites au sujet de deux de ces grandeurs.

Tandis que, dans les conditions particulières de l'exemple précédent (à savoir $r_m = 1$, $r_n = 0$), le coefficient de corrélation exprime un pourcentage, le carré de ce coefficient représente un pourcentage, d'une nature bien différente, il est vrai, dans une distribution homoscédastique (c'est-à-dire a dispersion constante). Désignons par σ_y l'écart type d'une des variables y, et par σ'_y l'écart type de la variable liée, c'est-à-dire l'écart type des valeurs de y correspondant à une valeur donnée de x. Alors

$$r^2 = 1 - \left(\frac{\sigma'_y}{\sigma_y}\right)^2 = \frac{\sigma_y^2 - \sigma'^2_y}{\sigma_y^2}$$

La carré du coefficient de corrélation montre donc quelle fraction de la dispersion d'une variable disparaît quand la valeur de l'autre variable a été fixée, ou, en d'autres termes, quelle partie des fluctuations d'une variable est déterminée par les fluctuations de l'autre variable. Cette propriété est assez connue et fréquemment appliquée, de sorte qu'il soit superflu de donner ici de nouveaux exemples.

Mais un autre schéma a permis d'établir des relations non moins simples mais rarement usées en pratique, et il vaudra la peine de les analyser plus en détail. Supposons que x et y sont des combinaisons linéaires de variables u_1, u_2, ...u_i, ...u_m et v_1 v_2, ... v_i, ... v_n dépendant les unes des autres, c'est-à-dire, les corrélations entre elles ne sont pas toutes nulles ($r_{u_i v_j} \neq 0$). Il n'est pas difficile d'écrire la formule générale pour le coefficient de corrélation entre

$$x = \sum_{i=1}^{m} a_i u_i \qquad \text{et} \qquad y = \sum_{i=1}^{n} b_i v_i$$

en fonction des paramètres a_i et b_i, des écarts types σ_{u_i} et σ_{v_i} et des corrélations $r_{u_i v_j}$. Or. ce n'est pas le cas général qui est intéressant, mais les formules concises qu'on obtient en adoptant différentes hypothèses simplificatrices.

Pour fixer les idées, nous désignerons par « simple » tout exemple où l'on suppose que les variables u_i (et de même les v_i) sont stochastiquement indépendantes les unes des autres, et seulement quelques-unes des variables u_i sont corrélées avec certaines variables v_i. Tous les cas particuliers qui ont été étudiés par d'autres auteurs appartiennent à la classe des exemples « simples »; nous nous bornerons donc ici à établir la formule générale et ne mentionnerons que brièvement quelques-unes des formules devenues classiques. Si l'on suppose que les u_i et les v_i ont des écarts de leurs moyennes respectives, l'expression générale du coefficient de corrélation entre x et y, pour les exemples « simples », sera

$$r = \frac{\sum_{i,j=1}^{i=m,\ j=n} a_i b_j \sigma_{u_i} \sigma_{v_j} r_{u_i v_j}}{\sqrt{\sum_{i=1}^{m} a_i^2 \sigma^2_{u_i} \cdot \sum_{j=1}^{n} b_j^2 \sigma^2_{v_j}}}$$

A partir de cette expression, il est facile de déduire la plupart des formules établies par différents auteurs.

a) Une formule donnée par Darmois (4) résulte des hypothèses suivantes :

$$\sigma_{u_1} = \sigma_{u_2} = \ldots = \sigma_{u_m} = \sigma_{v_1} = \sigma_{v_2} = \ldots = \sigma_{v_n};$$

$$r_{u_i v_j} = \begin{cases} 1 \text{ pour } i = j, \\ 0 \text{ pour } i \neq j; \end{cases} \qquad m = n,$$

Alors

$$r = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \cdot \sum_{i=1}^{n} b_i^2}}$$

Voilà les hypothèses qui sont à la base de la théorie des facteurs de la psychologie expérimentale. Cette théorie suppose que la corrélation entre la cote x obtenue par un individu dans un certain test et la cote y obtenue par le même individu dans un autre test provient du fait que les mêmes aptitudes $u_1, u_2 \ldots, u_n$ ont été utilisées dans une mesure a_1 respectivement $a_2 \ldots, a_n$ pour le premier test et dans une autre mesure b_1 respectivement $b_2 \ldots, b_n$ pour l'autre test. Ce n'est pas ici la place de discuter la théorie des facteurs mentaux; mais il est évident que la méthode est acceptable seulement dans le cas où, en effet, il existe des aptitudes mentales stochastiquement indépendantes les unes des autres.

b) Les hypothèses qui amènent à une formule établie par Yule (8) sont les suivantes :

$$m = n = 2; \quad r_{u_1 v_1} = 1; \quad r_{u_1 v_2} = r_{u_2 v} = r_{u_2 v_2} = 0; \quad a_1 = a_2 = b_1 = b_2 = 1.$$

Dans ce cas

$$r = \frac{\sigma_{u_1} \sigma_v}{\sqrt{\sigma^2 u_1 + \sigma^2 u_2} \cdot \sqrt{\sigma^2 v_1 + \sigma^2 v_2}} = \frac{\sigma_u}{\sigma_x} \frac{\sigma_{v_1}}{\sigma_y}.$$

'Le coefficient de corrélation est donc égal à la moyenne géométrique des rapports entre la dispersion (σ^2) de la partie liée (u_1 respectivement v_1) de la variable et la dispersion de la variable elle-même.

c) Dans un exemple donné par Croxton et Cowden (2), ces auteurs posent :

$$\sigma_u = \sigma_{u_1} = \ldots = \sigma_{u_m} = \sigma_{v_1} = \sigma_{v_2} = \ldots \sigma_{v_n}; \; a_i = b_i = 1;$$

$$r_{u_i v_j} = \begin{cases} 1 \text{ pour } i = j \leqq p, \\ 0 \text{ pour } i = j > p \text{ et } i \neq j. \end{cases}$$

Dans ces conditions

$$r = \frac{p}{\sqrt{m\, n}}$$

d) La formule de Bowley (1) n'est, en effet, qu'un cas particulier de l'exemple précédent, à savoir, celui où $m = n$. Alors

$$r = \frac{p}{m}.$$

Or, ici encore, le coefficient de corrélation, multiplié par 100, peut être considéré comme un pourcentage. Bowley lui-même a estimé que cette inter-

prétation de la valeur numérique de r est la plus simple qu'on puisse concevoir, en s'exprimant de la façon suivante : « Le coefficient de corrélation tend à être le rapport entre le nombre de causes communes à la création de deux variables et le nombre total de causes indépendantes desquelles les deux variables dépendent. » Le mot « cause » est évidemment mal choisi, comme a déjà constaté Keynes (6); mais l'idée qui est à la base de la conclusion de Bowley reste correcte. La formule de Bowley s'applique malheureusement presque uniquement aux jeux de hasard, car, en général, on n'est que rarement en présence de phénomènes qui dépendent de facteurs strictement indépendants les uns des autres. Le meilleur exemple est fourni par les expériences des dés de Darbishire (3), reproduites par Pearl (7). Darbishire a jeté n dés à la fois, mais, parmi ces n dés, p furent gardés sans altération d'un coup au suivant; le coefficient de corrélation observé entre le nombre de dés montrant un point pair dans deux coups consécutifs évidemment à peu près $\frac{p}{n}$.

Quant aux exemples qui n'entrent pas dans la catégorie désignée plus haut comme « simple », nous écrirons la formule générale du coefficient de corrélation seulement pour le cas où x et y sont composés de deux variables, c'est-à-dire $x = a_1 u_1 + a_2 u_2$ et $y = b_1 v_1 + b_2 v_2$.

Alors le coefficient de corrélation entre x et y est donné par

$$r = \frac{a_1 b_1 \sigma_{u_1} \sigma_{v_1} r_{u_1 v_1} + a_1 b_2 \sigma_{u_1} \sigma_{v_2} r_{u_1 v_2} + a_2 b_1 \sigma_{u_2} \sigma_{v_1} r_{u_2 v_1} + a_2 b_2 \sigma_{u_2} \sigma_{v_2} r_{u_2 v_2}}{\sqrt{a^2_1 \sigma^2_{u_1} + a^2_2 \sigma^2_{u_2} + 2 a_1 a_2 \sigma_{u_1} \sigma_{u_2} r_{u_1 u_2}} \cdot \sqrt{b^2_1 \sigma^2_{v_1} + b^2_2 \sigma^2_{v_2} + 2 b_1 b_2 \sigma_{v_1} \sigma_{v_2} r_{v_1 v_2}}}$$

Les formules « simples » peuvent être déduites immédiatement de cette dernière expression en posant $r_{u_1 u_2} = r_{v_1 v_2} = 0$.

Le cas particulier qui nous paraît le plus important est celui où $a_1 = b_1 = b_2 = 1$ et $a_2 = 0$, et $x = u_1 = v_1$.

Dans ce cas

$$r = \frac{\sigma_{v_1} + \sigma_{v_2} r_{v_1 v_2}}{\sqrt{\sigma^2_{v_1} + \sigma^2_{v_2} + 2 \sigma_{v_1} \sigma_{v_2} r_{v_1 v_2}}}.$$

On pourra interpréter ces conditions de la façon suivante : la variable y se compose de deux parties $v_1 = x$ et v_2. Si la corrélation entre la variable $(y = v_1 + v_2)$ et une partie de la variable (v_1) et les écarts types sont donnés, on peut calculer immédiatement la corrélation entre les deux parties de la variable $(r_{v_1 v_2})$. L'importance de cette relation du point de vue de l'interprétation de la valeur numérique de r repose dans le fait qu'elle permet de distinguer deux différents facteurs qui interviennent souvent dans la création d'une corrélation : l'un des facteurs est le fait que l'une des variables est une partie intégrante de l'autre variable et l'autre facteur est la corrélation qui existe entre la première variable et la différence entre les deux variables. L'application de la formule sera particulièrement fructueuse dans le domaine de la psychologie où les recherches se basent si souvent sur des corrélations, étant donné que des mesures absolues — et des différences entre les mesures — ne peuvent pas être obtenues. Si $v_1 = x$ sont les cotes obtenues par un groupe d'élèves dans un test au début d'une année scolaire et y les cotes obtenues dans un test correspondant à la fin de l'année, le coefficient de corrélation

entre x et y pourra être déterminé expérimentalement. Mais la corrélation entre la capacité des élèves au début de l'année et leurs progrès accomplis ne peut être déterminée directement. Or, cette corrélation importante sera estimée à l'aide de la dernière formule.

Une autre application de la formule générale peut être conçue de la façon suivante. Supposons qu'un pédagogue soit intéressé à trouver la corrélation entre le progrès fait par un groupe d'élèves en deux années consécutives. Évidemment, seules les corrélations entre le niveau atteint au début de l'expérience, à la fin de la première et à la fin de la deuxième année peuvent être déterminées. En désignant ces niveaux par u_1, u_2 et v_1, les progrès faits pendant les deux années seront $x = u_2 - u_1$ et $y = v_1 - u_2$, et la formule générale pour r pourra être appliquée, si l'on pose $v_2 = u_2$, $a_2 = b_1 = 1$ et $a_1 = b_2 = -1$. Dans ce cas, le coefficient de corrélation sera :

$$r = \frac{-\sigma_{u_1}\sigma_{v_1} r_{u_1 v_1} + \sigma_{u_1}\sigma_{u_2} r_{u_1 u_2} + \sigma_{u_2}\sigma_{v_1} r_{u_2 v_1} - \sigma^2 u_2}{\sqrt{\sigma^2 u_1 + \sigma^2 u_2 - 2\,\sigma_{u_1}\sigma_{u_2} r_{u_1 u_2}} \cdot \sqrt{\sigma^2 v_1 + \sigma^2 u_2 - 2\,\sigma_{v_1}\sigma_{u_2} r_{v_1 u_2}}}$$

Trois différents schémas nous ont permis de déduire des formules qui facilitent l'interprétation de la valeur numérique du coefficient de corrélation : le premier schéma se base sur l'idée que les paires de mesures appartiennent à de différents ensembles; dans le deuxième schéma nous nous sommes servis du concept de la dispersion de la variable liée, et les formules du troisième groupe ont été obtenues en considérant les mesures comme des combinaisons linéaires de grandeurs élémentaires liées entre elles. Évidemment, nous n'avions pas eu l'intention de donner des règles valables dans toutes les circonstances et suffisant à tous les besoins. Nous espérons, pourtant que les idées exposées ici aideront à dissiper certains doutes au sujet de la signification physique de la valeur numérique du coefficient de corrélation. D'autre part, certaines parmi les formules que nous avons établies ne permettront pas seulement d'avancer vers la solution du problème que nous nous sommes posé, mais elles prouvent, en même temps, l'utilité de la méthode des corrélations à l'analyse de certains phénomènes qui, par leur nature, sont inaccessibles à la recherche directe.

M.-V. Musham.

BIBLIOGRAPHIE

1. Bowley A.-L., *Elements of Statistics*, p. 355. London, 1926.
2. Croxton F.-E. and Cowden D.-J., *Applied General Statistics*, p. 664. New-York, 1945.
3. Darbishire A.-D. « Some Tables for Illustrating Statistical Correlation ». *Mem. and Proc. Manchester Lit. and Phil. Soc.*, vol. 51, 1907.
4. Darmois G., *Statistique et Applications*, p. 131. Paris, 1934.
5. Fisher R.-A., « The Resemblance between Twins, etc. » *Genetics*, vol. 10, 1925.
6. Keynes J.-M., *A Treatise on Probability*, p. 425. London, 1929.
7. Pearl R., *Introduction to Medical Biometry and Statistics*, p. 366. Philadelphia, 1930.
8. Yule G.-U., *An Introduction to the Theory of Statistics*, p. 227, London, 1910.